A Million Cancer Genome Warehouse

David Haussler
David A. Patterson
Mark Diekhans
Armando Fox
Michael Jordan
Anthony D. Joseph
Singer Ma
Benedict Paten
Scott Shenker
Taylor Sittler
Ion Stoica

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-211
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html

November 20, 2012

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
20 NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Million Cancer Genome Warehouse	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California at Berkeley,Electrical Engineering and Computer Sciences,Berkeley,CA,94720**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**61**	

Acknowledgement

This research is supported in part by the sponsors of the UC Berkeley AMP Lab: NSF CISE Expeditions award CCF-1139158, gifts from Amazon Web Services, Google, SAP, Blue Goji, Cisco, Cloudera, Ericsson, General Electric, Hewlett Packard, Huawei, Intel, Microsoft, NetApp, Oracle, Quanta, Splunk, VMware and by DARPA (contract #FA8650-11-C-7136).

A Million Cancer Genome Warehouse

David Haussler (UCSC), David Patterson (UCB), Mark Diekhans (UCSC),
Armando Fox (UCB), Michael Jordan (UCB), Anthony Joseph (UCB), Singer Ma (UCSC),
Benedict Paten (UCSC), Scott Shenker (UCB), Taylor Sittler (UCSF), and Ion Stoica (UCB)

UC Berkeley, UC San Francisco, and UC Santa Cruz

Draft 3.0
November 20, 2012

Executive Summary

Technology advances will soon enable us to sequence a person's genome for less than $1,000, which will lead to an exponential increase in the number of sequenced genomes. The potential of this advance is blunted unless this information is associated with patient clinical data, collected together, and made available in a form that researchers can use. Indeed, a recent US National Academy of Sciences study highlighted the creation of a large-scale information commons for biomedical research including DNA and related molecular information as a national priority in biomedicine, leading to a new era of "Precision Medicine." Based on the current trajectory, the genomic warehouse will be the heart of the information commons. To create it requires cooperation from a wide range of stakeholders and experts: patients, physicians, clinics, payers, biomedical researchers, computer scientists, and social scientists. Here we focus on the technological issues in building a genomic warehouse.

Genomic information is relevant to all aspects of medicine. To avoid a piecemeal approach, it makes sense to build a single database containing the inherited (germline) genome of millions of individuals that is used broadly and that is maintained throughout each individual's lifetime, rather than having a separate database of genomes for each specific project or disease, or for each specific disease episode in a patient's life. Further, in addition to a single germline genome, in cancer and immunology in particular, it is also essential to ascertain somatic mutations (disease causing mutations to DNA in cells that are not in the germline, and hence not passed on to children) to the genome that occur in particular cells during the lifetime of the patient. A cancer patient's data will soon include several full tumor genomes from biopsies at different sites and disease stages. This database would be augmented by RNA-sequencing data from specific tissues at specific times in a patient's life to capture gene expression, and by other molecular information, such as epigenetic changes (which do not swap DNA bases) and protein expression changes (modifications to the proteins that are produced from genes).

We focus on cancer in part because it is the most complex form of genetic data for a genome warehouse--setting a high water mark in terms of design requirements--but also because it represents the most acute need and opportunity in genome-based precision medicine today.

- It is widespread: one third of US women will face cancer in their lifetimes, one half of US men will also face cancer, and there 1.6M new cases in the US each year;
- Cancer is a disease that is clearly caused by changes to the genome;
- It is highly individualized, requiring precision medicine;
- Mortality is very high - one quarter of US deaths are due to cancer; within the next few years cancer will surpass heart disease as the leading killer;
- Speed is required - cancer patients have their genome(s) sequenced because they have a life-threatening disease and want genetic information that may inform treatment decisions that must be made in a matter of weeks, in contrast to individuals who are sequenced for the purposes of understanding long term predisposition to disease.
- There are an estimated 800 targeted therapies available or in the pipeline for cancer that may benefit from knowledge of the patient's tumor genome; and
- Its cost of care is much more than the $1,000 to sequence a genome.

As the Pulitzer Prize winning book *The Emperor of All Maladies* concludes, “The question is not if we will get this immortal disease, but when.” Given the numbers above, accelerating progress to discover effective cancer treatments could literally save millions of lives. Thus, cancer by itself is a compelling justification for a genome warehouse.

Within five years the collection of substantial amounts of molecular information from cancer patients, especially DNA, will be a part of patient care. Assuming appropriate privacy/legal protections, given a choice, most patients will want their information to also be available for research. They will understand that only by contributing to a large-scale effort can enough cases be assembled for study that are similar enough to their own subtype to draw meaningful scientific and medical conclusions. What we learn from their cancer will help their children and future generations with similar cancers.

To maximize use and facilitate interaction between research and clinical practice, the bulk of the data should be mirrored for dependability, but act as a single research database under one access infrastructure and as a clinical decision support database under another. These databases will have distinct Application Programming Interfaces according to the different purpose and compliance requirements, so that third parties can build both compliant diagnostic tests and research-oriented analysis software on top of the basic information commons. We need to adopt the best practices from information technology to build dependable tools for genetic analysis and to benchmark results so as to make the most rapid progress.

Whole genome sequences are not too much data to store. We show that it is now technically possible to reliably store and analyze 1 million genomes and related clinical and pathological data, which would match the demand for 2014. Moreover, thanks to advances in cloud computing, it is surprisingly affordable: multiple estimates agree on a technology cost of about $25 a year per genome. While the focus is on technology, to be thorough, this paper touches on high-level policy issues as well as low-level details about statistics and the price of computer memory to cover the scope of the issues that a million cancer genome warehouse raises.

There a number of concerns that this white paper addresses:

1. **Dual Use in the Clinic and in Research** - The only way to have a million genomes for research is to share data between research and clinical practice. We propose a single database with two different access structures and interfaces, one for clinical use and one for research. The database will provide one basic applications programming interface (API), and different third party entities will build either research or clinical service applications on top of that API on a competitive basis. A benefit of this is that sharing of clinical research data also facilitates the rapid deployment of research findings in the clinic.
2. **Unified Data and Computation** - Given that the cost-performance of network bandwidth inside a datacenter is about 100 times better than the Internet bandwidth outside the datacenter, it will be much more economical to compute within datacenters rather than across them. Moreover, cloud computing has demonstrated that purchasing,

housing, and operating tens of thousands of servers as a single "warehouse scale computer" yields economies of scale that offer a 5- to 10-fold reduction in the total cost of computation. To prevent data loss due to a catastrophe, we recommend two nearly identical datacenters in different geographic regions. While both will hold the same data, they could support different kinds of access, computation, or compliance infrastructure.

3. **Genomic analysis expense and accuracy** – Current popular software pipelines take days to turn the output of sequencing machines into genomes and genome comparisons, and the accuracy and biomedical depth of the analysis may not be good enough to enable precision medicine. The data processing portion of sequencing a genome currently costs about $1000, which is far too high. In addition, DNA of cancer tumors has much greater variation from standard reference sequences than the DNA of normal cells, which makes accurate analysis for cancer even more challenging. We need advances in algorithms and implementations to improve upon the current state-of-the-art in performance, accuracy, biomedical depth, and software-engineering practices to match the demand from an exponentially increasing number of cancer genomes. Moreover, we need to agree upon analysis metrics and benchmarks so that we can measure progress and thereby accelerate it.
4. **Single common patient consent form** – If every collection of 1000 genomes has its own consent form, which is the case today, then researchers would need to get permission from 1000 data access committee to use the 1M genome warehouse. Obviously, such an impediment would render the warehouse unusable. Aggregated cancer genomic data collected under a single consent form will be extraordinarily more valuable than data collected under many different consents.
5. **Need to further isolate some datasets** – Some data require special isolation. For example, data from patients in clinical trials must be tightly controlled until the study is complete. To handle this, we propose that a virtual cluster within the cloud computing facility, accessible only to those involved in the trial, be allocated to each ongoing trial that elects to use the genomics services provided by the database. In exchange for services, it would be expected that a significant portion of the data be opened up for general research after the trial is complete. A similar approach could be taken to some other types of studies, However, if researchers must get permission from an endless set of panels to access all the data, it will both slow progress and discourage those who could otherwise work on different data sets with fewer roadblocks. It is critical that we streamline and shorten the path from collection of data to research analysis on aggregated datasets in order that the discoveries made from this repository rapidly benefit the patients who need them.
6. **Regional warehouses and Application Programming Interfaces** – Many countries will not feel comfortable putting their citizens' health information into a single worldwide system that may be housed in and subject to the laws of a single nation. Hence, there will likely be multiple genomic warehouses, at minimum of one per continent. Third parties writing software for either clinical or research applications will need standards. Without planning, the likely outcome would be multiple incompatible systems. Following the successful tradition in the IT industry, we recommend identifying and embracing application programming interfaces early that will allow researchers and developers to

access to information in a more controlled manner.

7. **Privacy/use concerns** – While nearly everyone recognizes the desire for and importance of patient anonymity and/or appropriate limitations on data use, heavy-handed restrictions could slow progress and make it an unattractive set of data for researchers to work on. As long we follow common sense approaches to patient privacy, we can still help cancer patients in a timely manner. For example, within the US the HIPAA (Health Insurance Portability and Accountability Act) restrictions are routinely met for several cloud-housed datasets, so these potential obstacles are not as imposing as they might appear to outsiders. Getting the privacy and data use issue right is essential and requires a national dialogue. The National Cancer Institute (NCI) has suggested the possibility of a national program for Cancer Information Donors, the American Society for Clinical Oncology (ASCO) has proposed a rapid learning system for better clinical outcomes, and the American Association for Cancer Research (AACR) has provided a variety of information and communication resources for cancer researchers. These institutions, with input from patient advocates, could support a dialogue to address privacy and data use concerns, as well as help rapidly populate a 1M genomic repository.
8. **Data compression and statistical validity** – The purpose of building a genome warehouse is not just to store genomes, but also to enable a wide range of biological and medical inferences. Our calculations suggest that storing 100 gigabytes of compressed data per genome (100 petabytes for 1 million genomes) is the right compromise between cost and loss of information. As sequencing becomes cheaper and machines produce more data per genome, more aggressive data compression will be required to hold to the 100 gigabyte per genome upper limit. The consequence is that the data we keep will necessarily be incomplete and imperfect. Inference from cancer genome data will necessarily take place in a low signal/high noise regime, and it is essential to take this fact into account in designing data analysis pipelines. Building a statistically-aware genome warehouse raises a number of challenges. First, there are computational challenges involved in computing and propagating error bars on uncertain quantities. Second, database operations in the genomic warehouse need to be statistically grounded. Finally, to establish causality, the statistical underpinnings of causal inference (distinctions between experimental and observational data, attention to sampling frames, and so on) need to be provided in the genome warehouse.
9. **Embracing Technological Advances** – The dramatic reduction in sequencing costs is due in part to rapid innovation in sequencing equipment. The third generation of sequencing is currently being deployed, and the fourth generation is just around the corner. Thus, a repository must be neutral to the technology (and the manufacturer) that produces the sequencing information. It will be vital to the acceptance and success of a genome warehouse that it easily include all information sources. Whereas today some centers use custom assays of subsets of genes, with the dropping cost we suggest the minimum today should be whole exomes, and soon should be whole genomes. Moore's Law continues on the information technology side as well, so both the warehouse-scale computer design and the funding model must include a plan to refresh the IT equipment every few years, in part to grow the capacity of the warehouse.

10. **Potential Candidates** – We believe the first serious prototypes must be designed to scale quickly to many genomes, include patient records, and address privacy concerns. While a daunting combination, these necessary ingredients can be assembled. For example, some countries with national healthcare systems, such as Denmark (6M citizens), are already discussing plans to record germline genetic information of their citizens. For countries without national healthcare systems, like the US, we recommend pursuing multiple paths that collectively can develop the technology and make the case for the a single cancer repository:
    a. A *Top-Down* path via a large government run clinical trial to collect structured patient medical and genetic data.
    b. A *Patient-Push* path via a national organization that collects genetic and medical records directly from patients by leveraging patient advocacy groups.
    c. A *Center-Push* path forming a consortium of enlightened cancer centers to combine full cancer genomes and unstructured electronic medical records.
11. **Finding Seed Funding** – If a cancer genome warehouse proves to be as valuable in fighting cancer as we hope, its health benefits will be unquestioned. At these prices, we believe healthcare systems can easily pay its ongoing costs as it will save money and save lives. However, we need to find a path to bootstrap the funding for the first several years to give the genomic warehouse the chance to prove its worth. We make no recommendation as to whether the source should be government funds, philanthropic foundations, or industrial support. However, we strongly recommend that seed funds be found *soon* so that we can accelerate discovery of effective treatments and thereby more rapidly reduce the mortality of this voracious disease.

In establishing that it is technically and economically feasible to construct a million cancer genome warehouse, and describing the key characteristics that it must have, we hope to provide additional impetus for the many pioneers in biomedicine, both nationally and internationally, who are currently advocating such repositories. Like many other computer scientists, we are inventing new algorithms and building new tools that could help fulfill the vision of precision medicine, but we cannot make serious progress on them without access to real data at large scale. A team approach can change that. It is time to invest, and unleash the combined power of genomics and large-scale computation in medicine.

Introduction

This white paper discusses the motivation and issues surrounding the development of a repository and associated computational infrastructure to house and process a million genomes to help battle cancer, which we call the *Million Cancer Genome Warehouse*. It is proposed as an example of an information commons and a computing system that will bring about "precision medicine," coupling established clinical–pathological indexes with state-of-the-art molecular profiling to create diagnostic, prognostic, and therapeutic strategies precisely tailored to each patient's individual requirements[1].

The goal of the white paper is to stimulate discussion so as to help reach consensus about the need to construct a Million Cancer Genome Warehouse and what its nature should be. To try to anticipate concerns, including thorough cost estimates, it covers topics as varied as high-level health policy issues to low-level details about statistical analysis, data formats and structures, software design, and hardware construction and cost.

Problem Statement: Why a Million Cancer Genome Warehouse?

Cancer is a disease caused by combinations of mutations that occasionally accumulate in some cells in our bodies over our lifetime, and cause these cells to grow and divide in an uncontrolled fashion, often invading other tissues. In the United States, there are approximately 1.6 million new cases of cancer each year, 0.6 million die of cancer each year, and 13 million people live with cancer (about 4%) [2]. Although of enormous value to science[3],[4] and often relevant to treatment[5], the mutations that occur in newly diagnosed tumors are not being systematically assayed at this time. The cost of DNA sequencing is one impediment. However, this cost has dropped approximately 10,000-fold over the last 8 years, and in a year or so it will cost less than $1000 to sequence a human genome, making this assay feasible for routine clinical use.

Another impediment is our current difficulty in interpreting the DNA mutations we find. The only solution to improve the quality of these interpretations is more research, which requires a large dataset of human cancer genetic data. We believe we need a minimum of 1 million cancer genomes because smaller repositories will not have a sufficient number of samples for specific cancer subtypes to discover meaningful patterns. Without a large, aggregated database we lack statistical power.

---

[1] Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for Precision Medicine. *New England Journal of Medicine*.

[2] (2012). Cancer Facts & Figures - 2012 - American Cancer Society. Retrieved July 22, 2012, from http://www.cancer.org/Research/CancerFactsFigures/ACSPC-031941.

[3] Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, *456*(7218), 66-72.

[4] Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., et al. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993-998.

[5] Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., et al. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, *486*(7403), 353-360.

Following a recent US National Academy of Sciences study calling for a large scale knowledge network for biomedical research[6], Barbara Wold, former Director of Center for Cancer Genomics at the National Cancer Institute (NCI), suggested the possibility of a national program for Cancer Information Donors. These are cancer patients who elect to donate their cancer genomics data for research independently from their clinical treatment, much like making a blood donation. Such a program could enable the creation of a research database consisting of sequence data for 1 million cancer patients that could be coupled to a database for clinical use. It is one of several possible paths to this goal. Here we recommend a technical approach to building such a database. We hope that other authors will address the social and economic issues in creating such a database in greater depth in companion papers.

What data will be collected.

The use of high throughput sequencing data in clinical care has begun. Where tests for sequence mutations within specific exons of one gene were the norm, now targeted panels of genes are being simultaneously assayed. In cancer diagnostics particularly, the past several years have witnessed the development of targeted panels consisting of all of the exons for 20-300 genes[7]. Since the fundamental technique used to generate the sequence data is scalable (via use of additional PCR primers), the number and coverage of genes in these panels is expected to grow as the cost of sequencing continues to drop.

Although ultimately, all of these data will be collected systematically without specific targeting, there are several reasons why this has not yet happened:

- While the cost of sequencing continues to drop at an impressive rate, the cost of performing genome-wide sequence analysis in cancer is still too high for routine clinical use.
- Storing and interpreting smaller panels is a simpler and currently more tractable task.
- Since the fraction of the sample representing tumor can be small and the tumor population itself heterogeneous, it is desirable to sequence to a depth sufficient to accurately discern the sequence of a clone present in 1% of cells, in some cases less. Achieving this depth is much more tractable with targeted panels.
- Targeted panels include only sequence data from regions that can inform an immediate treatment decision.
- Development of targeted panels is in line with current practice of applying only CLIA-approved clinical tests evaluating a specific hypothesis or issue associated with a treatment decision in a situation where (1) the accuracy of the test is well-documented and (2) the proper use of the test results in making treatment decisions is well-documented.
- Collecting untargeted data may pose a liability issue for medical institutions and related companies, since disease associations are still being discovered. This opens up the possibility of a successful lawsuit in situations where mutations are recorded and

---

[6] Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for Precision Medicine. *NEJM.*

[7] Mamanova *et al* (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*. 7(2):111-8.

attributed to a specific patient even though they are not associated with a particular prognosis or treatment option at the time of diagnosis. This includes the majority of situations where the basis of the lawsuit may be an entirely unrelated act of claimed negligence.

A well designed patient consent form can and should help alleviate concerns about negligence. Additionally, there are distinct benefits to sequencing more broadly: namely the discovery of new disease associations and the ability to use new associations to help with existing clinical care and potentially help prevent recurrence or identify it early. These points are also addressed in several other sections of this paper.

The resource we propose within this paper will be designed to collect all types of data required to help diagnose cancer and guide its treatment, as well as information primarily useful for research. This repository can incorporate data commonly obtained today, such as array data from expression arrays or SNP chips and sequence from targeted panels of genes, but will be able to scale to RNA-seq and full genomic data, which is several orders of magnitude larger than these currently used datasets. Additional assays for epigenetic, non-coding RNA, metabolic, and protein expression data are providing important insights, will likely be used in clinical care, and are expected to be saved in this repository.

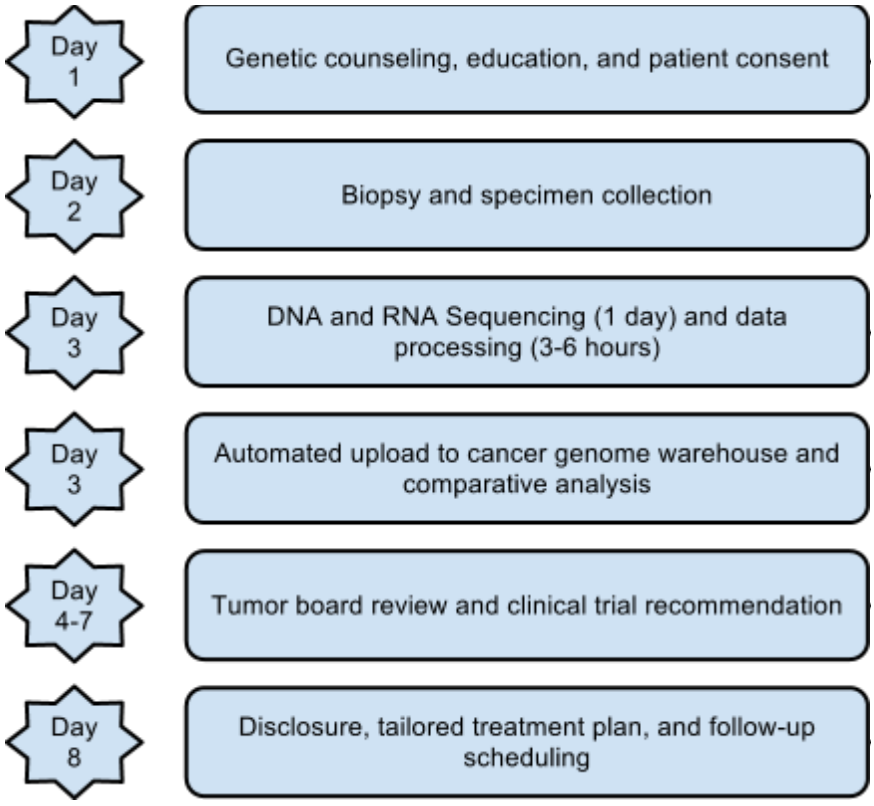
For the purposes of this paper, calculations will be performed using genomic data as an upper bound to the data storage requirements for each patient. We address the technical issues in scaling up. The policy, liability and other social issues mentioned above will have to be addressed separately.

How This Will Impact Patient Care

Despite the issues discussed above, several institutions have already adopted preliminary pipelines for interpretation of large genomic datasets within clinical oncology[8] and medical genetics[9]. The figure below shows an outline for a typical sequence of events following new clinical suspicion of neoplasm under the new molecular paradigm, drawn largely from existing processes published by these institutions. We expect the process to take roughly one week given a tumor board that meets biweekly, using a modern mutation calling pipeline and the genomic warehouse presented in this paper. The fourth step, "Automated upload to cancer genome warehouse and comparative analysis," is new to these processes.

[8] Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, et al. (2011) Personalized oncology through integrative high-throughput sequencing: a pilot study. *Science Translational Medicine*.

[9] Biesecker, Leslie G. (2010). Exome sequencing makes medical genomics a reality. *Nature Genetics*.



Current genomic testing strategies in oncology may combine tests for mutations in specific genes, low pass genome sequencing, exome sequencing (in which just the protein-coding regions are sequenced), RNA-seq and/or additional expression microarrays (in which RNA transcripts are sequenced or otherwise assayed to determine genes that are abnormally expressed), methylation/epigenetic analysis (in which modifications to chromatin by DNA methylation and other mechanisms are assayed by sequencing or microarrays) or SNP array testing (in which microarrays are used to initially determine tumor sample purity and to determine variations in the copy number of genes).

In the near future, as microarrays are replaced by sequencing and sequencing machines become network-attached devices, patient assays will be immediately and automatically streamed into large databases for storage and processing, either along with clinical information or accompanied by a URL indicating where clinical information can be retrieved. At this point, if a million cancer genome warehouse is created, patient data could be compared to a vast number of other cases that have been previously analyzed and contain treatment and outcome data associated with them. Such comparisons and analysis could provide the physician with valuable information about the specific mutations present in the patient's tumor that may be prognostic or suggest gene-targeted therapies, epigenetic changes and changes in gene expression and molecular pathway activity that define subtypes that are also prognostic or suggest particular therapies, and summary information about trials, studies and outcomes for similar cases found by matching patient data within the large repository.

Molecular analysis of cancer is vital to both clinical decision support and to research. Rendering this information into a form that is interpretable and actionable by physicians as well as researchers will be a very active area for third party software development and innovation. These third parties will rely on the Applications Development Interface and query platform provided by the centralized database. The section on "Performance Demands" addresses how a million cancer genome warehouse could achieve the speed and scale necessary to allow external programs to interpret genetic data for a patient within one day.

Most institutions engaged in clinical sequencing use tumor boards to review each case, to choose appropriate clinical trials for each patient, and to recommend tailored therapy. These boards generally have members from pathology, oncology, clinical genetics, and statistics or informatics. Additionally, some boards include ethics members; this can be particularly important when considering the implications of analysis and disclosure of such comprehensive genetic information. We envision that tumor boards will be the primary consumer of new molecular pathology information from the million cancer genome warehouse for clinical applications. By providing a single platform and standards on which third parties can build tools, interfaces and analysis pipelines for use by tumor boards across the country, the warehouse will foster a widespread use of molecular information in the treatment of cancer.

Tumor sequencing boards also play an important role in improving clinical care by identifying and incorporating new associations between genetic elements, therapies, and clinical outcomes generated in basic research. A dual-use million cancer genome warehouse could serve as a conduit between researchers and clinicians, helping translational researchers work with tumor boards to flag important new developments and decide when to incorporate them into patient care. This process has the potential to increase the pace at which scientific discoveries are translated into clinical practice. It will also accelerate changes expected in clinical trial design to meet the demands of personalized oncology[10].

Since this warehouse would be continually updated, it could be used by clinicians to identify in their patients newly discovered clinical correlations from other studies. It would provide a single source for the myriad specialists involved in a patient's care to access relevant molecular data and to be alerted to potential new disease correlations or treatments involving these molecular data that are available to their existing patients. Given the complexities of today's cancer regimens and the number of clinicians involved, this continuously enhanced information could improve the quality and continuity of care[11].

Indeed, this post hoc analysis of full genomes and resulting change to treatment is one argument why patients should have their full tumor genomes sequenced even though today we only a fraction of it to prescribe treatments. A patient who has their cancer genome sequenced today, even if it has no effect on their immediate treatment, will be in a better position if they recur 10 years from now because it will help doctors in the future to understand the origins and

---

[10] Maitland, Michael L. and Schilsky, Richard L. (2011) Clinical Trials in the Era of Personalized Oncology. *CA: A Cancer Journal for Clinicians*.

[11] www.asco.org/CancerLinQ

long term progression of their disease.

How it will Impact Science and Discovery of New Treatments

Even though they were running blind with respect to the full spectrum of cancer mutations, cancer researchers over the last 40 years have made remarkable progress in understanding the disease, and along the way have revolutionized our basic understanding of the signaling and other molecular pathways of the cell[12]. This advance occurred through intensive experimental investigation comparing different kinds of cancer-mutated cells to normal cells. The models that have been built are greatly oversimplified, but with a million genome database, and computer-enabled deeper and more comprehensive models[13],[14],[15] the scientific community would now be in a dramatically better position to begin to tackle the actual complexity of the cellular processes that drive life.

As the most powerful and prolific source of naturally mutated human cells for study *in vitro*, via xenograft and *in vivo* within patients, cancer studies will continue to play a key role in advancing basic molecular, cell and developmental biology. A one million cancer genome repository with an effective API will provide an essential support structure for this work, akin to the way Google's search engine and the World Wide Web support so many other critical aspects of society today.

The science that results from a deeper molecular understanding of normal and cancer cells has already had a profound impact on cancer treatment. New targeted therapies include imatinib[16] (Gleevec) inhibiting the receptor tyrosine kinase ABL, trastuzumab[17] (Herceptin) against

---

[12] Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646-674.

[13] Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, *26*(12), i237-i245.

[14] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., & Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature genetics*, *37*(4), 382-390.

[15] Chuang, H., Lee, E., Liu, Y., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, *3*(1).

[16] Druker, B. J., Guilhot, F., O'Brien, S. G., Gathmann, I., Kantarjian, H., Gattermann, N., et al. (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *New England Journal of Medicine*, *355*(23), 2408-2417.

[17] Slamon, D., Leyland-Jones, B., Shak, S., Paton, V., Bajamonde, A., Fleming, T., et al. (1998). Addition of Herceptin [TM](humanized anti-HER2 antibody) to first line chemotherapy for HER2 overexpressing metastatic breast cancer (HER2+/MBC) markedly increases anticancer activity: a randomized, multinational controlled phase III trial. *Proc.Annual Meeting-American Society Of Clinical Oncology. American Society Of Clinical Oncology*.

ERBB2, gefitinib[18] and erlotinib[19] against EGFR, crizotinib[20] against ALK, vemurafenib[21] against BRAF, as well as various inhibitors of genes in the PI3K/AKT/mTOR pathway and other key cancer drivers. These drugs show dramatic results when the mutations they target are present in the tumor, with vastly more precision than standard chemo- or radiation therapy. Genomics analysis is also emerging as a key player in understanding differential patient responses to both targeted and standard therapies, and emergence of resistance.

Cancer genomics studies are now pushing discovery at a greatly accelerated rate. New processes have been implicated in tumorigenesis such as mutations in splicing factors[22] [23], and in HLA genes[24], as well as translocations in epithelial tumors[25]. New aberrant metabolic processes have been linked to glioblastoma and other cancers via the discovery of mutations of genes such as IDH1[26]. Mutations found in chromatin modifying genes have revealed the key role of epigenetics in many cancers, including childhood neuroblastoma[27], renal carcinoma[28], multiple myeloma[29], AML[30] and many others. (IDH1 mutation in fact has a dual

---

[18] Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, *350*(21), 2129-2139.

[19] Shepherd, F. A., Rodrigues Pereira, J., Ciuleanu, T., Tan, E. H., Hirsh, V., Thongprasert, S., et al. (2005). Erlotinib in previously treated non–small-cell lung cancer. *New England Journal of Medicine*, *353*(2), 123-132.

[20] Gandhi, L., & Jänne, P. A. (2012). Crizotinib for ALK-Rearranged Non–Small Cell Lung Cancer: A New Targeted Therapy for a New Target. *Clinical Cancer Research*, *18*(14), 3737-3742.

[21] Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., et al. (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine*.

[22] Yoshida, Kenichi et al. "Frequent pathway mutations of splicing machinery in myelodysplasia." *Nature* 478.7367 (2011): 64-69.

[23] Graubert, Timothy A et al. "Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes." *Nature genetics* (2011).

[24] Comprehensive genomic characterization of squamous cell lung cancers, The Cancer Genome Atlas Research Network, *Nature*, in press.

[25] Tomlins, Scott A et al. "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer." *Science* 310.5748 (2005): 644-648.

[26] Parsons, D Williams et al. "An integrated genomic analysis of human glioblastoma multiforme." *Science's STKE* 321.5897 (2008): 1807.

[27] Cheung, Nai-Kong V et al. "Association of age at diagnosis and genetic mutations in patients with neuroblastoma." *JAMA: the journal of the American Medical Association* 307.10 (2012): 1062-1071.

[28] Varela, Ignacio et al. "Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma." *Nature* 469.7331 (2011): 539-542.

[29] Van Haaften, Gijs et al. "Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer." *Nature genetics* 41.5 (2009): 521-523.

[30] Ley, Timothy J et al. "DNMT3A mutations in acute myeloid leukemia." *New England Journal of Medicine* 363.25 (2010): 2424-2433.

and perhaps more important role in epigenetics[31]). New cancer mutational mechanisms such as chromothripsis[32] and kataegis[33] have been discovered, along with hypermutated subtypes of colorectal and endometrial cancers with distinct clinical features. Researchers have found striking, extensive subclonal evolution[34] and mutational heterogeneity in breast[35], kidney[36] and other cancers.

About 800 drugs are targeted at about 300 different genes relevant to cancer are either available or in development[37]. We desperately need better molecular characterization of patients and better tracking of outcomes in order to target these therapies appropriately. Thus, the genome warehouse will not only be an engine of discovery for targeted therapies, but a vital part of clinical translation, validation, and refinement of the cancer pathway knowledge it fosters.

Beyond single-agent therapeutics lies the new horizon of combination therapies[38], synthetic lethal and passenger mutation-based strategies[39], immunotherapies[40], and other molecular strategies. The need for a comprehensive molecular and clinical database will only increase as the science of cancer pushes forward, and as precision treatments that anticipate and adapt to mutational and epigenetic moves made by the cancer become the new standard of care.

Current Repositories, Data Format, Data Size

Repositories that currently hold personal genome sequences include the dbGaP database of genomes at the US National Center for Biotechnology Information (NCBI), the European Genome-phenome Archive (EGA), the DNA Databank of Japan (DDBJ), and the NCI Cancer Genomics Hub (CGHub). Reference genomes and other genome datasets are also available from the Beijing Genomics Institute (BGI) and its journal GigaScience. Each of these centers makes the genomes they contain available for research.

---

[31] Figueroa, Maria E et al. "Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation." *Cancer cell* 18.6 (2010): 553-567.

[32] Stephens, Philip J et al. "Massive genomic rearrangement acquired in a single catastrophic event during cancer development." *Cell* 144.1 (2011): 27-40.

[33] Nik-Zainal, Serena et al. "Mutational processes molding the genomes of 21 breast cancers." *Cell* (2012).

[34] Shah, Sohrab P et al. "The clonal and mutational evolution spectrum of primary triple-negative breast cancers." *Nature* 486.7403 (2012): 395-399.

[35] The genetic determinants of the molecula portraitrs of human breast tumors, The Cancer Genome Atlas Network, to appear.

[36] TCGA, in preparation.

[37] http://www.businessweek.com/magazine/content/11_26/b4234024330707.htm

[38] Al-Lazikani, B., Banerji, U., & Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nature Biotechnology*, *30*(7), 679-692.

[39] Muller, F. L., Colla, S., Aquilanti, E., Manzo, V. E., Genovese, G., Lee, J., et al. (2012). Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature*, *488*(7411), 337-342.

[40] Rosenberg, S. A. (2005). Cancer immunotherapy comes of age. *Nature clinical practice. Oncology*, *2*(3), 115.

A common data format for DNA sequences is used by all centers: the BAM format pioneered by the international 1000 Genomes Project[41]. A BAM file stores individual snippets of DNA a few hundred DNA bases in length, read from random positions in the genomes of the cells from the tissue sample[42]. To get sufficiently accurate information, each position in the tumor genome may be read on average 30 or more times[43]. For cancer, a minimum of two BAM files are obtained for each patient, one for the DNA of the normal cells (representing the germline genome of the patient), and one for DNA of the cells of the tumor tissue. In practice, RNA is often read as well, and provides valuable information on which genes in the tumor are abnormally expressed. RNA information is also stored in the BAM format.

BAM is also used for "exome" sequencing files, which only cover the protein coding region of the genome and a few other selected regions. For reasons given above, today many centers assay a subset of genes needed for treatment of a particular cancer. As costs drop, the number of genes is expanded. Given the rapid change in costs and that it takes time to agree on which subset of genes to sequence, we think today it makes more sense that the absolute minimum sequencing should be whole exome. Similarly, because the price of whole genome sequencing is dropping faster than that of special preparations used in exome sequencing, it is likely that most assays will move to cover whole genomes.

The total size of the BAM files of a single patient may approach 1 terabyte (TB) or 1 million TB for 1 million patients. As of July 2012, there were approximately 1400 TB of raw BAM data in NCBI, 1100 TB in EGA, 312 TB in CGHub, and approximately 35 TB of research data at BGI/ GigaScience. These numbers do not include RAID replication or backup copies. DDBJ has an exchange agreement with NCBI and EGA for unrestricted data and thus will have a similar set of BAM files. One million TB is considerably larger than current databases.

A million TB is 1,000 petabytes (PB), which is roughly the size of all the videos in YouTube. One can expect in the near future to gain a factor of approximately 10 from future BAM compression technologies such as CRAM[44] or cSRA[45], with very little loss of medically or scientifically relevant information. Therefore we discuss the design of a 100 petabyte database to hold genomics data for 1 million cancer cases.

Limitations of Data Compression

Considerable research has gone into video compression, while BAM compression is currently in its infancy. One might argue that it would be wiser to spend money on better compression

---

[41] Altshuler, D. M., Lander, E. S., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., et al. (2010). A map of human genome variation from population scale sequencing. *Nature*, *467*(7319), 1061-1073.
[42] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.
[43] Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, *11*(10), 685-696.
[44] Fritz, M. H., Leinonen, R., Cochrane, G., & Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome research*, *21*(5), 734-740.
[45] http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

methods now rather than building a large genome warehouse. However, significant effort has gone into the CRAM and cSRA schemes as well, and while a factor of 10 is feasible it will not be possible to compress BAM format DNA sequence data by a factor of, say, 1,000 without losing very important information. The reason is rather subtle. Cancer tissues contain a population of cells with slightly different mutations, multiplying at different rates, and occupying different proportions of the tumor tissue[46],[47]. Mutations that occur in 1% or fewer of the cells in the tumor tissue may be important in determining how the tumor will respond to therapy that destroys most of the other cells, as well as in predicting the future course of the tumor without treatment[48],[49]. Yet even reading each position in the genome 100 times, a novel important mutation in 1% of the cells at a particular position in the genome cannot be distinguished with sufficiently high accuracy from the enormous pool of background noise within the reads of the BAM file (see the Error Characteristics Table).

Two years after a patient's tumor genome is sequenced, a pattern may be discovered that indicates the importance of a mutation at a particular location in the genome for a particular type of cancer. Only then can we go back and recognize that among the vast pool of randomly discordant reads, those few discordant reads in the original patient mapping to this position were indicating that important mutation in a small subset of the cells in his tumor. If we only keep a record of the obvious mutations supported by a large percentage of the reads, we can't go back to check. Thus, to retain all the medically and scientifically useful information from the sequencing of a tumor, we must retain all the essential information from each of the cancer DNA reads. Once again, if we do keep all the information, the patient who does full genome sequencing can benefit from later discoveries, as centers could contact the patient and treating physician about the impact of the discoveries on treatment.

API for Research at Different Data Summary Levels

Not all research requires the full 100 petabytes of DNA reads. Therefore, we propose a layered structure, in which the compressed BAM files are kept in the slowest and cheapest medium, while index, summary and interpretive information is kept on faster but more expensive media. In addition to basic biomedical research, we expect the database will be used by biotech and biopharmaceutical companies to help design new therapies, diagnostic companies to design new tests, epidemiologists to study mutation prevalence, and most importantly by physicians using third-party clinical decision support software. We must build an Applications Programming Interface (API) to the database that allows research scientists and engineers in both the public and private sector to build software tools that help interpret the data and translate it for use in research and in clinical practice. In typical operating mode, many of these tools will only

---

[46] Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, *486*(7403), 395-399.

[47] Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., et al. (2012). The Life History of 21 Breast Cancers. *Cell*.

[48] Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*.

[49] Downing, J. R., Wilson, R. K., Zhang, J., Mardis, E. R., Pui, C., Ding, L., et al. (2012). The Pediatric Cancer Genome Project. *Nature Genetics*, *44*(6), 619-622.

need access to intermediate interpretations or high-level summaries of the data, which are 10s to 1000s of times smaller than the compressed BAM files. The clinical data associated with each patient falls into this latter category of relatively small, high-level data. Uniformity and completeness of clinical information, however, remains a serious issue[50],[51]. Current medical records in either structured form or as unstructured text are trivial in size when compared to the BAM data, but must be further regularized to allow effective large-scale computational analysis. As this effort cuts across all of medicine, we would hope that the cancer genome database effort could link-up with an even broader effort, rather than solve this problem in isolation.

Need to Bring the Computation to the Data

It is not yet feasible to transmit petabytes of data over optical fiber from one datacenter to another in different parts of the country. Our national backbone fiber operates at 10 Gigabits per second, at most half of which would be available for any particular transfer, which translates into 50 terabytes per day, or 20 days per petabyte. CGHub, for example, has achieved these rates. Wider use will drive this rate down sharply, because no single user would get half of the capacity of the national backbone fiber all to themselves. Even after the data are successfully transferred, there are considerable costs associated with storage at the other end. Therefore, the only practical way forward for a database of this size is to bring the computation that needs to be done to the database, rather than transmitting the data to a remote site for computation. Computation involving in-depth comparison of large numbers of genomes cannot be done efficiently without transferring the data if genomes are spread over many sites. This means that to the greatest extent possible, the genome data should all be located in a single database and that attached to that database flexible “cloud” computing resources should be provided, which many laboratories can use either on a temporary or permanent basis. This is a key element of our design.

Queries of a Million Cancer Genome Warehouse

At a simplified level, mutations affect genes, genes are organized into pathways of molecular activity in the cell, and disruptions in these activities cause disease. Therefore, some of the questions that researchers and clinicians will be asking using information in the database are:

- For a given subtype of cancer:
    - What are the pathway disruptions that are characteristically present in the tumor cells?
    - Are there specific genes involved in those pathways that are recurrently mutated in this or other cancer subtypes to create these disruptions?
    - Are there specific mutations in these genes that recur in many cases and are common *driver mutations* for this or other subtypes of cancer?
    - Is there a genetic signature associated with prognosis, response to therapy, or drug efficacy and side effects for that subtype of cancer?

---

[50] Bennett, S. N., Caporaso, N., Fitzpatrick, A. L., Agrawal, A., Barnes, K., Boyd, H. A., et al. (2011). Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genetic epidemiology*, *35*(3), 159-173.

[51] Covitz, P. (2004). National Cancer Institute External Data Standards Review.

- For a given type of pathway disruption:
    - In what subtypes of cancer does it often occur?
    - What are the phenotypic/clinical consequences and how are such disruptions assayed?
    - What genes in the pathway are typically disrupted?
    - What therapies are available to counteract or exploit this pathway disruption, e.g. by inhibiting the action of these genes?

- For a given gene:
    - In what pathways and cancer subtypes is it involved?
    - What biology is known about the gene, e.g. what is its normal variation within the human population and its role in specific pathways?
    - What are the differential effects of distinct mutations on this gene, e.g. loss-of-function mutations versus various types of gain-of-function mutations?

- For a tumor tissue sample from a given patient:
    - What are the specific mutations, gene changes, and pathway disruptions relative to other samples from that patient, i.e. relative to the patient's normal tissue sample or to other samples taken during the course of the diagnosis, treatment and recurrence?
    - How common are these mutations?

- For each individual cancer patient:
    - How does the set of mutations in a patient's genome compare to mutations observed in other similar patients? How were those patients treated and what was the outcome?
    - What genetic signatures can be identified for that patient that are statistically associated with prognosis, response to therapy, drug efficacy, or side effects?
    - What controlled studies relevant to the patient are discussed in the literature, and for what on-going clinical trials might the patient qualify?

The goal of the database is to provide a platform and an API upon which third-party groups could build software services that answer questions such as these. It is a non-trivial problem to organize 100 petabytes of personal DNA information to enable efficient and secure queries of these kinds.

Database Architecture

It is essential to have a layered architecture for a database of this size supporting a large variety of queries. Figure 1 shows the three-layered approach that we propose, where each level is optimized for the data attributes, size, and queries associated with the given data types.

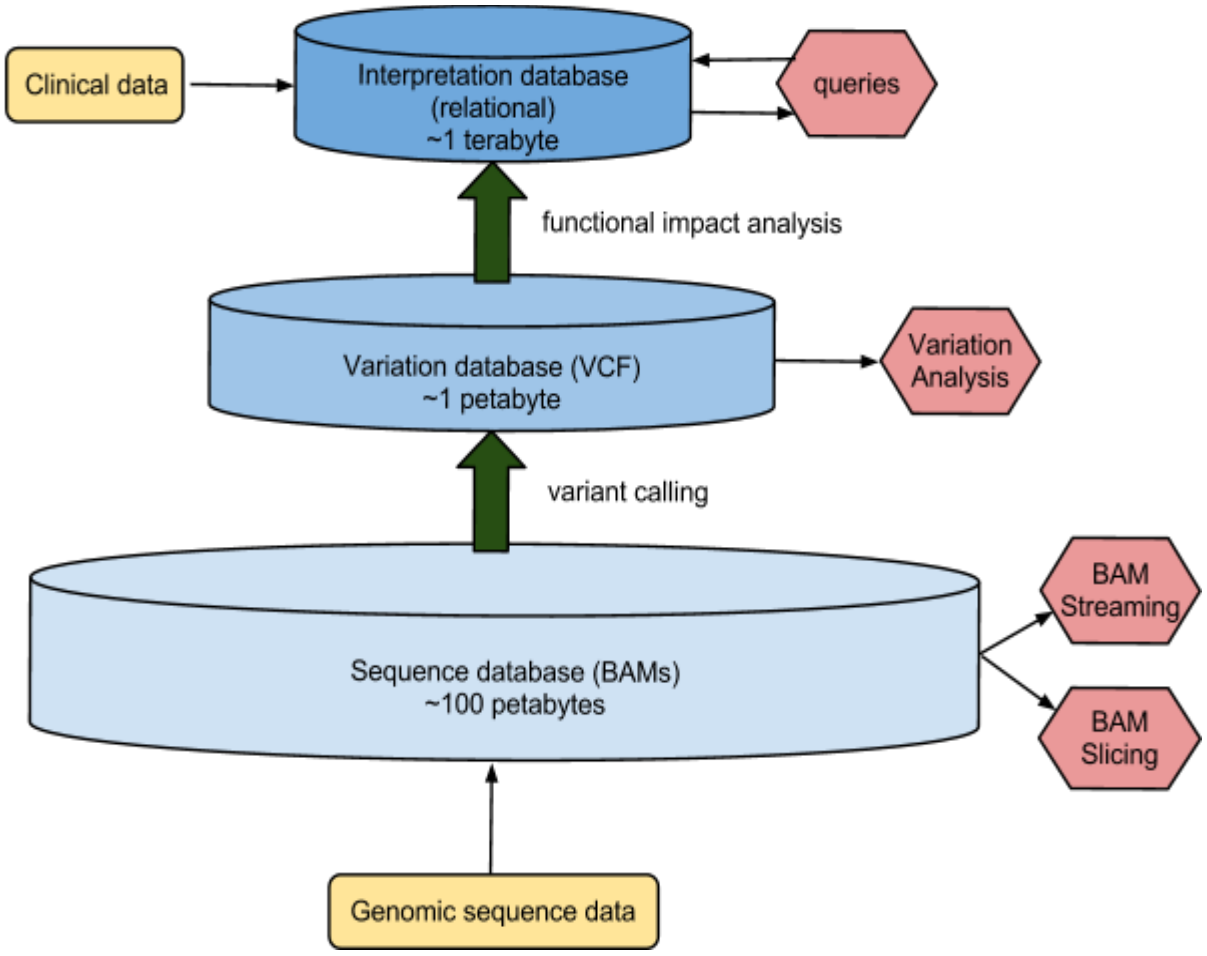


Figure 1: Million Cancer Genomes Database Architecture

Focusing on the DNA and RNA data only, for the moment, the lowest layer is the storage of the compressed BAM files. Bulk access to these will be less frequent than derived data. However, the database must support relatively rapid access to specific information in them. In particular, all DNA reads in a BAM file are mapped to specific coordinates in the current universal reference human genome determined by the Genome Reference Consortium[52] (see Appendix A). Rapid, random access queries of the form "give me all DNA reads for a given sample that map within a certain narrow range of locations in the reference genome" must be possible without downloading entire BAM files. This query is known as *BAM slicing*. At a deeper level, this layer should support compression and queries based on identity by descent of large segments of genomes referred to as *haplotypes* (Appendix A). Multiplexing accesses to the different types of data in BAM files (e.g. bases vs quality scores) will reduce access times, as discussed below. In addition to these optimizations for specific queries, it will still be necessary to be able to relatively rapidly stream all the data from 1000s of genomes into a research analysis pipeline running in the nearby servers for in-depth analysis.

[52] (2008). Genome Reference Consortium (GRC) - NCBI - NIH. Retrieved July 22, 2012, from http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.

The next layer is called the variation database or VCF layer after the increasingly popular Variant Call Format for genome data[53],[54],[55]. A VCF file for a patient is constructed from all of the BAM files for the DNA reads from the different tissue samples taken during the course of the patient's lifetime. Thus, in moving from the sequence database layer up to the variation database layer, the data are reorganized in a patient-centric manner, grouped according to longitudinal information about a single individual. The VCF file does not record the raw DNA reads, but only records what differences are inferred from the BAM files to be present in the genome(s) in each tissue sample relative to a standard genome. The standard is typically the universal reference human genome, but for cancer could in the future instead be the genome inferred to be present from birth in the patient's normal cells, called the patient's *germline genome*.

Because a VCF file records only the differences between closely related genomes, including *inherited variants* that distinguish the patient's germline genome from the universal reference genome and *somatic mutations* that distinguish the patient's tumor cells from his normal cells, it is typically more than 100 times smaller than the BAM files from which it is derived. The price of this compression is that some evidence for specific mutations that are present at a very low level is omitted from the VCF file. Currently, the optimal conversion of BAM to VCF is still a difficult data interpretation problem and is a very active area of research. However, we expect this process to be hardened by the time this database is built, with robust software in place that has well-defined statistical accuracy. Special data structures representing the tumor genomes at different times in the patient's life that interpret the single nucleotide inherited and somatic variants as well as the larger structural changes represented in a VCF file for an individual will also be an integral part of this layer (Appendix A). To obtain the maximal accuracy, all genomes obtained from an individual should be processed together. This means that we cannot just produce a new VCF file and freeze it each time a new BAM file for the patient arrives from the sequencing pipeline, but rather all previous VCF interpretations of the patient's past recorded mutations must be reevaluated, even though research has been conducted and medical actions may have already been taken in light of those previous interpretations.

Built upon the VCF layer is the interpretation database layer, which is the first layer in which phenotypic and clinical data are incorporated. As discussed in the NAS Precision Medicine Report[56], the interpretation database must be primarily organized around the individual patient, with tables that can be joined using a unique patient identifier as a key. The database layer will take advantage of patient-centric organization already present in the VCF layer. Other joins

---

[53] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

[54] (2011). VCF (Variant Call Format) version 4.1 | 1000 Genomes. Retrieved July 22, 2012, from http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41.

[55] (2012). PyVCF Download - Softpedia. Retrieved July 22, 2012, from http://linux.softpedia.com/progDownload/PyVCF-Download-79117.html.

[56] Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for Precision Medicine. *New England Journal of Medicine*.

must be supported as well, such as based on sample ID within a patient, and by phenotypic/ clinical feature, reference genome location (and common ethnic variant haplotype for that location), mutation, gene, pathway, and disease. For efficiency, the interpretation database layer can only contain summaries of the data present at the lower levels. Mechanisms for data update, revision, retraction, versioning, etc. must be coordinated between the interpretation level and the lower levels. Finally, the database layer must contain reasonably complete provenance information on how the data in the lower layers were obtained, including details of experimental protocols and algorithms used.

The interpretation data layer is the most heterogeneous from a software design perspective, and because our understanding of cancer is evolving quickly, it must be designed such that third parties can both contribute to it and build on top of it. Reasoning about the biology of cancer and its clinical outcomes involves sophisticated models of disparate phenomena ranging from molecular signalling pathways within the cell to global physiological response. The database must provide the tools that the research community needs to build and deploy such models, including or in close collaboration with a platform by which such models can be deployed in research and clinical decision making. The process of building the models themselves that answer the user questions listed above must not be dominated by the operators of the database. Rather, it must engage the widest possible set of scientists and engineers.

Performance Demands for a Million Cancer Genome Warehouse

Having covered the potential medical benefits of a warehouse, we need to learn its cost. Hence, we next dive down into the details of the computing needed to analyze the such data. (Readers less concerned with such low-level details may want to skip to the next section.)

Given the projected uses of this system, we expect a mix of production and research workloads. The production queries will be run for every patient at the time of diagnosis and then at follow up sequencing appointments as the patient's status is monitored. The research queries will be much more ad hoc, designed to support researchers in their exploration of these large datasets. The content of the research queries could vary widely. We expect research projects will probably require their own dedicated compute and block storage resources which will be provisioned by the sponsoring party.

Query Type	Query Size	Query Frequency
Production Diagnosis	< 1PB	1.6M / year
Production Check-up	< 1PB	13M / 6 mo = 26M / year
Clinical Trial Research	10 TB - 1 PB (10 - 300 patients per trial with contextual info)	~8000 possible trials * 5 researchers / trial * 10 queries/researcher/day = 400K / year
Ad hoc research	variable (up to 100 PB)	100 simultaneous

Production Workload

Production workload will consist of a set of predefined queries. This workload will be generated by physicians for each new and existing cancer patient. Given that there are roughly 1.6M people diagnosed each year[57], we should plan for an average daily production load of 4,500 queries of this type. Additionally, there are roughly 13M people living with cancer and we will periodically want to perform a checkup every six months, meaning we will have to run on average 70,000 times daily. Rounding up, let us assume that we need to run 80,000 queries per day, or 1 query per second. To achieve such throughput we would likely need to store the query input data in fast storage (e.g., DRAM) and replicate it.

Assuming a query takes 30 seconds on average, we need to execute roughly 30 queries in parallel. Furthermore, assume that the input of each query is 1 petabyte and that each query executes in parallel on 1,000 cores, and the data is partitioned so that each core reads roughly 1 terabyte of data. Next, assume that each server has at least 16 cores, and that the level of parallelism on each server is 10x (we assume that the loss of 4x accounts for other tasks and overhead). In other words, every byte of input can be shared by 10 queries. If each query's input is 1 PB, we will need to replicate the input 3 times (as each 10 queries can share the same input). As a result, we need 3 PB DRAM and 30,000 cores to run 80,000 queries/day with an average query response time of 30 sec. Note that assuming 16-32 cores per server we need 1000-2000 servers to satisfy the computation demands. However, even assuming each server had 512 GB DRAM, we need as many as 6000 servers to satisfy the memory demands!

If the DRAM requirements are too expensive, another option would be to group (batch) all queries in a given time interval and run them together. By streaming the input data through all queries in the group, we need to read the input only *once* per batch of queries. Similarly, assume each batch of queries runs on 1000 cores, and that the input is on the disk. Since each core needs to read 1 TB, it will take a batch 5.6 hours to read the entire input in parallel, assuming that the disk throughput is 50 TB/sec. Accounting for contention and other overheads, let us assume that it takes 12 hours to read 1TB of data from a single disk. If we assume that the computation time of a query is still 1min, then we can batch 720 queries together. As a result, by trading the latency and waiting 12 hours we can achieve the query throughput target of 50K queries/year without requiring large amounts of DRAM.

Clinical Trial Research and other Special Studies

Part of the research workload is generated by large research institutes and hospitals (e.g., clinical trials). Ideally, the research workload should not interfere with production workload. The research workload is more explorative (ad hoc) in nature and will typically process significantly more data. Furthermore, some workloads will process the original BAM files. A likely approach to achieve isolation would be to allocate virtual clusters for each project. Let's assume that a special study may operate on ~100 TB of data, that is 1,000 patients at 100 GB/patient or 100 patients at 1 TB/patient. To initiate a research project we would dynamically spin off a virtual

[57] (2012). Cancer Facts & Figures - 2012 - American Cancer Society. Retrieved July 22, 2012, from http://www.cancer.org/Research/CancerFactsFigures/ACSPC-031941.

cluster and then copy the dataset to the virtual cluster. Assuming we stripe data over 100 disks, it will take around 6 hours to copy 100 TB of data, if each disk can sustain a throughput of 50 MB/second. Assuming 10 disks per server, we would need around 10 servers for each research project. In other words, we allocate 1 server per 10 TB of dataset. Note that we assume that these research projects will be long lived--days or even months--so that the time it takes to spin out a virtual cluster and copy the dataset is relatively small.

One important aspect of clinical trial research is the calculation of efficacy of treatment or identification of biomarkers predicting outcomes. These are highly statistical processes, generally requiring iterative processing steps most efficiently performed in DRAM. Since each trial will have a virtual cluster with DRAM and compute apportioned for the endeavor, these resources will be constantly available to perform these operations and allow trial researchers to monitor the progress and potential discoveries arising from their trial on a continuous basis. For a 100 patient trial, each of 10 servers contains 192-512GB DRAM, or about 20-50GB DRAM per patient.

Ad Hoc Research

Finally, we expect researchers will want to run queries on the raw data (i.e., BAM files). Most of the time, researchers will be interested in identifying either frequency of specific sequences within the data (looking for novel sequence, pathogens, frequency of a defined sequence, copy number variation, tumor or allele fraction, and so on) or recalculating occurrence of sequence variants either on an individual or population basis. The main challenge with these queries will be the disk bottleneck as the raw data is stored on the disk. Assuming the BAM files are stored on 5TB drives, and assuming again that the disk sequential read throughput is 50 MB/second it will take 28 hours just to read the data from the disk! Thus, the only realistic way to run ad-hoc queries that each require a large fraction of the raw data is to batch them. Assuming that a response time of one week is acceptable, we can answer virtually any ad hoc query by using 17% (= 28h/7days*24h) of the disk I/O. Similarly, we assume that these queries will also use up to 15-20% of the CPU resources. We believe this is a conservative assumption as we expect these queries to be I/O bounded.

This contrasts with ad hoc queries that require only a small fraction of the reads in the database, such as a BAM slicing query that pulls from multiple BAMs only the reads that map to a specific location in the reference genome. These types of queries will be frequent. For them, fast indexing by location in the reference genome and/or by specific DNA sequence can be used to avoid the I/O bottleneck described above.

Finally, since raw reads and quality scores are expected to make up the bulk of the data stored on the system, while many of the downstream analysis will be interested only in frequency or variant data, routine batching should occur in one coordinated, weekly update to compute and store this summary information. Caching this data could tremendously reduce the size of most queries if it eliminates loading of the raw data.

The following table summarizes the resource requirements and the response times for various query types.

Query Type	Query Size	Load	Resources	Response Time
Production (diagnosis & checkup)	1 PB	27.5M queries/year	30K cores (1000 - 2000 servers); 3PB DRAM (6000 servers @ 512GB/ server)	30s with 3 PB DRAM, 12h otherwise (reading from disk)
Clinical Trial Research / Special Study	10TB - 10PB	N/A	1-100 servers / study	N/A
Ad hoc research	variable (up to 100PB)	100 simultaneous	15-20% of entire cluster	1 week

Software Principles for a Million Cancer Genome Warehouse

The hardware described above is a necessary condition to build a warehouse, but its usefulness is limited by the quality and extensibility of the software that runs in the warehouse. Hence, it is vital that such software be built using the best practices from the information technology industry. Moreover, to ensure that the efforts in inventing new software algorithms for genetic analysis are headed in the right direction, we need to agree on benchmarks to measure progress.

Service Oriented Architecture

A major element of the success of "Web 2.0" has been a strong and rapid trend towards *service-oriented architecture (SOA)*[58]. In this design stance, an application is structured as a collection of independent but cooperating services, and access to data and functionality occurs exclusively through each service's Application Programming Interface (API). The distinction is subtle but important: if service A needs data from service B, it *must* call an existing API on B to get that data, because no other access is provided to access B's database directly. Applications are then constructed by composing the desired services. For example, the retail site Amazon.com composes standalone services for retail catalog search, recommendations, order fulfillment, creating product reviews, and integrating third-party marketplaces.

The advantages of SOA have established it as the de facto architecture for network-enabled applications:

- APIs tend to be narrow interfaces, so services can be evolved independently as long as the APIs remain stable. Services can be reengineered at the software level or moved to different hardware, perhaps radically so; for example, algorithms can be reimplemented to take advantage of emerging compute architectures such as Graphics Processing

[58] Fox, A. and Patterson, D. *Engineering Long-Lasting Software*, Strawberry Canyon, 2012.

Units (GPUs) without any changes needed to the client applications.

- Service-level APIs are designed to operate over a wide-area network protocol. While in practice the cooperating services may be physically co-located, APIs designed this way are necessarily language-neutral and language-independent. The rapid rate of progress in programming tools is a strong motivator to avoid creating a system that is heavily dependent on the use of any one language, however popular.
- When an application is composed of independent services, each service's resource needs can be addressed separately, with more resources being dedicated to compute-intensive or storage-intensive tasks in order to "balance the pipeline".
- As new algorithms and techniques are developed, in a SOA new applications can be created to use them by recomposition.
- The software infrastructure building blocks needed to create SOA—Web servers, network-based programming frameworks, testing frameworks, and so on—are not only mature but the focus of constant ongoing attention because of the move to SOA in the ecosystem of "commodity" Internet services.
- If a hardware or software "appliance" approach is desired, it is straightforward to create user interfaces in front of a composition of services. Indeed, some of the highest-volume Web services such as Twitter are architected just this way, so that the Twitter.com site is actually a relatively simple User Interface (UI) in front of a set of independent services.
- If new data formats emerge that are better suited for particular tasks within the research agenda, they can be transparently deployed as long as the APIs for working with the data remain stable.

All of these advantages argue for an SOA-based rather than full-application-based or appliance-based approach.

API Design Principles

Some design principles for these APIs include:

- APIs should be language-neutral, as we expect that software writers will use many different tools for the job. An easy way to achieve language-neutrality is to adopt widely-used SOA conventions such as REST (Representational State Transfer) for establishing the set of operations, HTTP (Hypertext Transfer Protocol) as the medium over which service calls occur, URLs (Universal Resource Locators) or UUIDs (Universally Unique Identifiers) for referring to data items, and simple intermediate representations such as XML (eXtensible Markup Language) and JSON (JavaScript Object Notation) for representing the calls to a service and the responses.
- APIs must be designed around the reality that the computation should be near the data. In general, API calls will name a set of data items, an operation to be performed, and a place to put the result. For example, Amazon's Elastic MapReduce API names datasets stored in an HDFS file system, which uses URLs as its naming scheme.
- Any specialized hardware (wet lab or otherwise) should be virtualized behind an appropriate API. It is less important to obtain wide agreement on how a particular class of device should be represented, as long as the API documentation and communication standards are open.

Scientific Software Engineering

Another software issue is the quality of the data analysis software itself. A 2008 survey of 2000 scientists found that most are most are self-taught in programming, that only a third think formal training in software engineering is important and, perhaps as a result, less than half have a good understanding of software testing.[59] Merali chronicles the widespread concern about the quality of scientific software, and reports of a case where a bug in a program supplied by another research lab forced a "structural-biology group led by Geoffrey Chang of the Scripps Research Institute … to retract five papers published in *Science*, the *Journal of Molecular Biology* and P*roceedings of the National Academy of Sciences*."[60]

Through trial and error, the IT industry has developed software development practices that make it easier to build dependable software that is easy to use and easy to install.[61] Best practices include using source code control systems; test driven development where the tests are written *before* the code; highly productive modern programming environments such as Ruby on Rails or Scala and Spark; "scrum" organization of small programming teams; calculating "velocity" to predict time to develop new features; and Agile software development lifecycle that involves rapid iterations of working prototypes and close interaction with end users for feedback.

Benchmarks

Finally, the IT industry made rapid progress in improving performance not only because of the additional hardware resources provided by Moore's Law, but also because the industry agreed on what were the proper metrics to measure performance and in particular what were the best benchmarks to run to have a fair comparison between competing systems.[62] Before that agreement, each company invented its own metrics and ran its own set of programs as benchmarks, making the results incomparable and customers suspicious of them. Even worse, engineers at competing companies couldn't tell whether innovations at their competitors were useful or not, so the competition occurred only *within* companies rather than between them. Once the IT industry agreed on a fair playing field, progress accelerated as engineers could see which ideas worked well and which didn't.[63]

Rapid progress in genomics will be problematic without proper metrics and benchmarks for alignment, assembly, SNP and structural variant calling, gene expression level and isoform interpretation from RNA-seq data, pathway analysis, and clinical outcome prediction.

---

[59] Hannay, J. E. *et al.* How do scientists develop and use scientific software? *Proc. 2nd Int. Workshop on Software Engineering for Computational Science and Engineering* (2009).

[60] Merali, Z. Computational science: ...Error Why scientific programming does not compute. *Nature* 467(2010), pp. 775-777.

[61] Fox, A. and Patterson, D. *Engineering Long-Lasting Software*, Strawberry Canyon, 2012.

[62] Patterson, D. For better or worse, benchmarks shape a field: technical perspective, *Communications of the ACM*, v.55 n.7, July 2012.

[63] Hennessy, J. and Patterson, D. *Computer Architecture: A Quantitative Approach*, 5th edition, Morgan Kaufmann Publisher, 2011.

In summary, open and language-neutral APIs using existing technologies will not only allow different applications to be created through the composition of services, but will prevent data format heterogeneity, hardware heterogeneity and software (language) heterogeneity from becoming obstacles to software progress. Even with good APIs, however, if analysis software development does not follow best development methodologies, it will take longer to build and be more difficult to use, which will slow progress.That is, the software engineering practices of genomic tool developers affect the speed that we can address the looming data analysis bottleneck. Finally, even if we agree on APIs and adopt the best software engineering practices, if we cannot agree on how to evaluate results, advances in data analysis software will not come as fast as we need; it is hard to make progress when you can't measure it.

Regional Warehouses and APIs

The laws that protect the privacy of citizens of a country where the warehouse is located normally do not protect the citizens of other countries. For example, US terrorism laws explicitly give US agencies the right to uncover private information about potential suspects in other countries. Thus, there will likely be at minimum of one warehouse per continent. Given this political reality, we imagine that researchers will first develop and validate hypotheses within their closest repository, and subsequently will want to test it in the others, and developers will first bring products to market in their own country, and subsequently want to adapt them for use in other countries. APIs that allow researchers to access to information in a more controlled manner will greatly facilitate the globalization of the new molecular approach to cancer.

Design and Cost to Build and Operate a Million Cancer Genome Warehouse

While we have covered the hardware requirements to analyze the data and offered guidelines on how to build useful software for the warehouse, to complete the cost estimate we need to expand the design to include the cost of reliably storing the genomes as well as the operational costs of a Million Cancer Genome Warehouse. (We again dive down into low-level details for technology aficianados, and readers with other interests may want to skip ahead.)

The Open Compute Project (OCP) developed standard designs for compute servers and storage servers to be used in datacenters. Initiated by Facebook in 2011, the goal is to reduce costs and improve energy efficiency by utilizing "vanity free" servers. The project now includes more than a dozen companies--including AMD, Dell, HP, Intel, and Quanta--and it recently released the second generation of designs. We expect to see new designs each year.

The OCP philosophy is to separate functions rather than have everything integrated into a single box that acts as a common building block. Hence

- There is both a storage-oriented server design and a compute-oriented server design, rather than have the same amount of processors, DRAM, and disks in every server.
- Power supplies are included in the cabinet, rather than included in each server.
- Batteries are in a separate rack, rather than in every rack.

To deliver 100 PB, we need 125 PB of raw storage capacity to handle multiple disk failures safely (RAID 7 with 3 check disks per 12 data disks). As we project 5 TB per 3.5 inch SATA disk in 2014, we need 25,000 disks. The storage-oriented OCP server, called Open Vault, contains 30 disks. A single "Open Rack" can contain 18 Open Vaults, yielding 540 disks or 2.7 PB per rack in 2014 (Each rack is approximately 21 inches wide and 90 inches tall.) Thus, we need about 45 racks for 125 PB.

For the compute server, we selected the Intel v2.0 motherboard, which uses 2 Intel® Xeon® E5-2600 processors, has 16 DIMM slots per board, and 4 PCIe slots (from x16 to x4). In 2014 we could build 192 GB per board and 10 cores per processor. Two motherboards fit in a chassis, with three chassis per tray, and 24 trays per Open Rack. Thus, one rack contains 144 compute-servers, of 288 processors, 2880 cores, and 27 TB of DRAM. If we went with 35 racks for computation to match the 35 racks for storage, the total would be approximately 10,000 processors, 100,000 cores, and 1 PB of DRAM.

In addition to single Open Racks, OCP offers a triple rack that has room for 3 Ethernet switches along with space for 24 compute-nodes or storage-nodes. If we add racks for batteries and datacenter switches, the total is about 85 racks of equipment.

CAPEX/OPEX of a Proposed Platform

The raw computer equipment cost in 2014 is roughly $2M for processors, $4M for disks, $5M for DRAM, $2M for networking equipment, and $1M for power supplies, chassis, racks, cabling, assembly, testing, and so on. Hence, the computing equipment capital expenditure (CAPEX) is approximately $14M per datacenter. With a second datacenter to prevent against data loss due to catastrophes, the total computing equipment CAPEX would be $28M.

(While this estimate may be missing some items, we don't include volume discounts nor buying over time as demand merits, which would surely save money at this scale of purchase.[64])

If we use the UC San Diego Supercomputing Center as a reference as an example colocation site, leaving it up to the facility to recover the costs for space, and cooling for 85 racks of equipment, charging scheme for rack space and cooling at $14k per rack, we would expect an annual OPEX of about $1.25M per datacenter.

The peak power demand for the equipment would be roughly 1 megawatt for processors and 1 megawatt for DRAM, disks, switches, inefficiencies due to power supplies, and so on. If the datacenter runs at a Power Usage Effectiveness[65] of 1.25, that would yield a peak demand of 2.5 megawatts. At $0.08 per kilowatt hour (in San Diego) and if we assume it operates on

[64]Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M. A view of cloud computing, *Communications of the ACM*, v.53 n.4, April 2010.
[65] Hoelzle, U., Barroso, L.A. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, Morgan and Claypool Publishers, 2009.

average of 70% of peak load, the annual electricity cost would be about $1.25M per datacenter.

The number of administrators to monitor and replace broken equipment would for such a facility typically be 6 to 12, so let us add another $2.5M for personnel (including benefits) per datacenter.

(While we could be missing some costs, note that it might well be possible to locate the datacenters in locations that would be cheaper than San Diego.)

The annual networking costs can be separated into costs for uploading data to the database and costs for downloading data from the database. The download depends on the demand, but we could expect the costs to be $5,000 per gigabit/second per month, and if most computation is done at the database, the average demand to be less than 4 gigabit/second. At $5K per Gbps/month, this costs $0.25M/year.

We expect an upload model similar to that used today by storage companies like Amazon, which involves direct upload of data via the network for the vast majority of clients, and physical shipping of disks for a select few extremely large producers. One such large producer might be a company that may emerge which specializes in fast and inexpensive whole genome sequencing for research purposes, and when approved as a clinical test, as a clinical service. The roadmap for improved network speeds calls for increase from 10 gigabit/s to 100 gigabits/s in the next few years, with projections for 100-400 gigabits/s by 2015 and 400-1,000 gigabits/s by 2020[66]. Therefore we do not see upload speed as a major limitation for the majority of lower volume upload clients. We allocate another $2M for networking costs for data upload.

Thus, our estimate for the two datacenters is roughly an initial $28M CAPEX and an initial annual OPEX of $8M. If we round up to account for omissions in these rough calculations, we end up with a $30M CAPEX and a $15M annual OPEX. If we upgrade all the computers, disks, and networking switches every 3 years, we could spread the CAPEX over 3 years ($10M per year), so the continuous OPEX would be roughly $25M for the hardware for both datacenters. Appendix C has an alternative design that reaches the same cost estimate.

Commercial Cloud Design

Another potentially interesting calculation is using the storage system of Amazon Web Services, called S3 for Simple Storage Service, to estimate costs. The current standard price for a small amount data is $125 per terabyte (TB) per month. The standard discounted price for at least 5000 TB or 5 petabytes (PB) is $55 per TB per month, or $55,000 per PB per month.

Since S3 storage is replicated internally, there is no need to account for additional redundant storage. Amazon S3 redundantly stores your objects on multiple devices across multiple facilities. For example, Amazon S3 is designed to sustain the concurrent loss of data in two facilities. Amazon claims that the chances of losing data are remote.

[66] http://www.ieee802.org/3/ad_hoc/bwa/BWA_Report.pdf

Suppose we could shrink the storage requirements from 2*125 PB as assumed above to "just" 40 PB. The cost would be $55,000/PB/month * 12 months * 40 PB = $26,400,000, or about $25 per genome per year. While this estimate is just the storage cost and does not include the computing costs:

- This estimate is based on 2012 prices, which are likely to drop by 2014.
- We only need pay for storage *occupied* versus storage *reserved* (as in the earlier estimates), so the average storage costs may be lower.

Such a design would also give the greatest flexibility in storage and computation, as there would be no fixed maximum or minimum computation or storage as in the prior designs. As discussed further in the section on data compression below, it may be challenging to achieve compression to 40 PB, but even at 80 PB, the proportional costs of $50 per genome per year may still be attractive given the other advantages of this approach, and the scale and social importance of this project may merit further discounts that would be negotiated with commercial cloud provider like Amazon Web Services.

Privacy Policies for a Million Cancer Genome Warehouse

The privacy of patient information in the context of healthcare and clinical research has traditionally been protected by two major federal laws: the Health Insurance Portability and Accountability Act (HIPAA) and the federal "Common Law" which outlines protection for human subjects involved in research. Over the past 10 years, "genetic privacy" policies for the specific protection of genetic information have been implemented in most states, in federal law (e.g. GINA), and in policies written by several of the National Institutes of Health. There are two primary - and often contentious - mandates that are addressed by these policies

- The protection of personal genetic information which can be used to identify an individual, his or her origin, family lineage or race, and personal traits or disease susceptibilities he or she or immediate family members may have
- The promotion of research involving genetic information to improve human health

There is an important balance between respect for individual privacy and access to specific data required for research to progress. Because of the personal and predictive nature of genetic information, its disclosure could influence an employer or potential insurer, or could otherwise divulge specific information an individual might not want to share. Conversely, access to large corpuses of such information unquestionably lends the statistical power required to make important disease associations. Particularly in the field of cancer, where alteration of genetic information is synonymous with disease, access to large and specific bodies of this data is necessary to identify trends or mechanisms of disease which lead to new therapies and diagnostic tests.

These two mandates are not always in conflict, however. In fact, the nature of genetic alterations present in cancer expose a subtle but important point of agreement between these mandates. While the unique combination of genetic variation (germline mutation) encoded in a person's genome is private, the alteration in that genome (somatic mutation) due to a neoplastic process is not. Somatic mutations are not passed on, do not identify established traits, and do

not identify an individual in the same way germline mutations do. Much research also relies on identification, aggregation and evaluation of somatic mutations and their association with clinical outcomes. Disclosure of this information to cancer researchers is highly useful in research and it is much less private. Genetic privacy policies largely allow the sharing of somatic mutations from patient records that are otherwise de-identified.

There have been several different approaches to striking the balance between the mandate for privacy and that for research[67]. Some groups have argued for open consent[68], the release of information by patients into the public domain for the study of cancer. This has been promoted by several groups, including the Personal Genome Project[69] and is starting to take hold. The concept of a "cancer information donor" introduced by the NCI recently is more conservative in that patient data would still remain under restricted access only to qualified researchers. A similar consent is contemplated by the "Portable Legal Consent"[70] used by SAGE Bionetworks[71] a non-profit organization promoting data intensive research and models in biomedicine. How to best manage consent remains an active topic of discussion[72].

One of the most difficult issues surrounding the creation of large databases of patient genome data for research is the question of how results obtained from the study of patient genomes by third parties other than the patient's own physician are returned either to the physician or directly to the patient at their election. Expectation of returning all possible results and associations to the patient may be dangerous since disease associations are tentative and clinical relevance is still being established, while returning too few results may not sufficiently serve the data-donating patient. To create a million cancer genome warehouse for research, it may be best to start with the blood donor metaphor, in which there is no expectation of direct benefit to the donor, and allow infrastructure for return of results for the benefit of the donor to grow organically on top of the database API.

We want to emphasize that the first step towards a useful cancer genome warehouse is a universal patient consent form within a region. The data collected before such consent is used will be much less useful than the data collected after such consent is used.

Appendix D describes technical approaches that may help address the privacy concerns for germline DNA data.

---

[67] Schadt, Eric E. "The changing privacy landscape in the era of big data." *Molecular Systems Biology* 8.1 (2012).

[68] Lunshof, J. E., Chadwick, R., Vorhaus, D. B., & Church, G. M. (2008). From genetic privacy to open consent. *Nature Reviews Genetics*, *9*(5), 406-411.

[69] Church, G. M. (2005). The personal genome project. *Molecular Systems Biology*, *1*(1).

[70] (2011). consent to research. Retrieved July 22, 2012, from http://weconsent.us/.

[71] (2009). Sage Bionetworks Seattle | Home. Retrieved July 22, 2012, from http://sagebase.org/.

[72] (2012). Informed consent: A broken contract : Nature News & Comment. Retrieved July 22, 2012, from http://www.nature.com/uidfinder/10.1038/486312a.

Structured vs. Unstructured Electronic Patient Records

It would seem that before trying to collect a million cancer genomes and the corresponding electronic patient records, we must first agree on standards for patient records so that we get a consistent dataset from which to draw inferences. While a certainly a good project to do, it could take years to reach agreement on what should be in the standard.

The Big Data movement is about unstructured information; indeed, many argue that what people mean by Big Data is not so much that it is large in size, but that it is unstructured, uncurated, and inconsistent data.[73] The Big Data challenge then is to mine information gold nuggets from the ore of dirty data, and many researchers and companies are building new algorithms and tools with exactly those goals. Engineers at Google argue in their paper "The Unreasonable Effectiveness of Data" that based on their successes at image analysis and natural language translation

> *"The first lesson of Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available. … invariably, simple models and a lot of data trump more elaborate models based on less data."*[74]

In the era of electronic medical records, data being collected in clinical data warehouses is beginning to be used to inform therapy decisions[75] as well as to perform surveillance on approved medications.[76] [77] In an example employing unstructured data, LePendu et al used term extraction tools on the clinical notes of a million patients to compile a database of statistically significant patterns of off-label drug use.[78] Preliminary results include a cancer drug being used to treat blockage of veins from the retina and a sleep disorder drug used to treat Parkinson's disease. They use such discoveries to identify adverse off-label usages that can then be incorporated into clinical practice. Most importantly, these examples illustrate the use of unstructured data to derive statistically significant, actionable associations from these heterogeneous datasets. Although still at an early stage, similar natural language processing techniques have been used to construct databases of formal relationships between terms within

---

[73] Patterson, D. Computer Scientists May Have What It Takes to Help Cure Cancer, *New York Times*. December 5, 2011.

[74] Halevy, A., Norvig, P. & Pereira, F. The Unreasonable Effectiveness of Data, *IEEE Intelligent Systems,* vol. 24, no. 2, 2009.

[75] Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the emr era. N. Engl. J. Med. 2011;365:1758-1759

[76] Lependu P, Iyer SV, Fairon C, Shah NH. Annotation analysis for testing drug safety signals using unstructured clinical notes. J Biomed Semantics. 2012;3 Suppl 1:S5

[77] Liu Y, Lependu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Summits Transl Sci Proc. 2012;2012:47-56

[78]LePendu P, et al. AMIA Analyzing Patterns of Drug Use in Clinical Notes for Patient Safety. Summit on Clinical Research Informatics. San Francisco, CA: 2012.

unstructured data that are being applied to help personalize therapy[79].

Given the urgency of advancement of cancer, it would seem wise to try the unstructured path in parallel with structured approaches, as the unstructured path will surely get to large scale much more quickly and allow new algorithms and tools to be created much sooner. Algorithmic insights that come about from the recent surge in Big Data research may well allow advances in treatment for some cancers years sooner than it would take to first standardize patient records and then get a million cancer genomes that record patient information in that standard.

Permission for Full Access to Data in the Million Cancer Genome Warehouse

Clearly, the power of the warehouse is the ability to cross traditional research boundaries to make discoveries that can accelerate improvements in health care for the benefit of society that ultimately is funding the research and the Million Cancer Genome Warehouse. While summary data will be available in an unrestricted manner as appropriate, privacy concerns are likely to force regulation of full access. Researchers needing full access will have to agree to terms set by patient consent and interpreted by an appropriate Data Access Committee.

Requiring separate permissions from panels responsible for each source of data does not scale to the Million Cancer Genome Warehouse. If this problem cannot be solved with a relatively simple permission mechanism, there is no point in building the warehouse, as it will be too painful for researchers to do their work. There are both procedural and technical examples in place today that could serve as templates for these permissions. Procedurally, the data access requirements for TCGA have specific usage and publishing clauses which could be adopted for this warehouse. From a technical perspective, cloud service companies such as Amazon and Google have developed robust and reliable solutions for partitioning access to machines for access by individual clients which could easily be adopted here. Ultimately, leading government agencies such as the National Cancer Institute (NCI), the American Association for Cancer Research (AACR), and the American Society of Clinical Oncology (ASCO), informed by patient advocates, must come up with a workable solution to this issue.

Potential Funding Models for a Million Cancer Genome Warehouse

Funding such a large endeavor will require investment from several different parties. We expect that interested parties include biomedical technology companies, pharmaceutical companies, information technology companies, the government (i.e. NCI and NIH), hospitals, research institutions, philanthropists, and eventually payers (medicare and private insurance). We see two phases:

1. **Startup Phase**, $30M for CAPEX + $15M/year of OPEX or $25M/year at a cloud provider.

[79] Whirl-Carrillo M, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012 Oct;92(4):414-7.

Initially, this will require large investment to procure the hardware to implement the project and pay the architects who will create this resource.

2. **Production Phase**, $25M/year of OPEX (including tech renewal) or $25M/year at a cloud provider.

   Once this process becomes established and used in clinical care, Medicare and private insurance would pay the bulk of the costs required to sustain this effort, considering it as part of the cost of a clinical test. Given the potential benefits to the patient, these storage costs and access for downstream care can reasonably be included in the price of the assay. One-time assays today such as Oncotype DX and Mammaprint are reimbursed at $2,000-$3,000 per test. Similar pricing should cover both the cost of assay generation and storage for any one of exome/transcriptome/methylome assay, storage, and re-analysis by 2014 using the production system outlined above. If the cost of genome sequencing continues to drop at the current pace, full genome sequencing can also be performed, stored, and periodically re-analyzed on this system within these cost constraints.

It is not clear how the start-up phase should be funded. Government funds channeled into typical government contracts do not seem likely to produce the desired result for a number of reasons that we will not elaborate on here. A true partnership in which the creators of the database are thought leaders, are respected and have creative control would be more likely to succeed. This may be possible with sufficient philanthropic leverage or through the establishment of a consortium model like the Hap-Map project[80], representing a collaborative public-private effort.

Storage Compression/Management

With the rapid growth of sequence data, efficient sequence compression methods have assumed a new importance. The two predominant methods are CRAM[81]; and cSRA[82], a component of the SRA Sequence Toolkit[83]. Both methods apply reference-based compression of mapped reads. Rather than storing the sequence of the reads, they store data describing how the reads differ from the reference genome: where substitutions, inserts, and deletions occur, and any bases that are substituted or inserted. Both methods also apply similar heuristics to discard secondary mappings, reduce the metadata stored per read, and store only those flags that cannot be recomputed accurately from compressed data. Both achieve lossless compression ratios on the order of 0.3 bits/byte.

---

[80] Gibbs, Richard A et al. "The international HapMap project." *Nature* 426.6968 (2003): 789-796.

[81] Fritz, M. H., Leinonen, R., Cochrane, G., & Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome research*, *21*(5), 734-740.

[82] http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

[83] Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic acids research*, *39*(suppl 1), D19-D21.

However, most of the contents by volume of FASTQ and BAM sequence files are not sequence data but per-base quality scores. These scores are typically discrete values ranging from 33 to 126, stored as one character per base. Compression performance will ultimately depend on reducing the storage of quality scores without significantly reducing the scientific content of the sequence archive. In lossless mode, both CRAM and cSRA retain the quality score verbatim. In lossy mode, cSRA discretizes the quality score, while CRAM applies a "quality budget" in which only the most important quality scores are preserved (these are typically the scores at substitutions or indels). Manufacturers are taking steps to reduce the number of bits required for lossless storage of quality values. For example, the Illumina sequencing equipment produces an 8 bit "quality score", which is considerably bigger than needed for reasonable accuracy and alignment algorithms. Steps they are taking to reduce the quality score to 4 or 5 bits will benefit lossless compression. Additional performance gains might be achieved by exploiting properties of quality scores, such as how the values are non-uniformly distributed with most appearing rarely while a few appear frequently; and how adjacent bases tend to have similar quality scores[84].

Sequencing capacity has far outstripped Moore's Law. Over the last ten years, sequencing capabilities have been increasing at more than four times the speed of computer processors[85]. This underscores the importance of dealing with compressed sequence data efficiently. Along those lines, the world of genomics could draw on some valuable lessons from the world of video compression. Part of the success of the MPEG standard was due to its usability, its ability to exploit network access, random access, and multiplexing. Each of these attributes apply to sequence data:

- *Usability*: It is essential to have a single API with plugin decompression algorithms. This will allow the client to tune the decompression scheme according to its precise needs and parameters.
- *Network access*: Just as one can now watch streaming video on demand, the sequence data should not need to live on the same physical machine as its client. The client software (and the decompression software) should be able to retrieve the sequence data over the network, invisibly to the user.
- *Random access*: One can jump into the middle of a streaming video. Likewise, the ability to index into the middle of a sequence archive is essential for methods that analyze the set of sequences that map to one given genomic region.
- *Multiplexing*: Video contains separate channels for visual and audio data. Likewise, sequence data could be composed of separate channels for mapped reads, quality scores, and other elements. This would give the client the ability to greatly improve bandwidth by disconnecting from any channel that is not needed.

[84] Loh, P., Baym, M., & Berger, B. (2012). Compressive genomics. *Nature Biotechnology*, *30*(7), 627-630.
[85] Wan, R., Anh, V. N., & Asai, K. (2012). Transformations for the compression of FASTQ quality scores of next-generation sequencing data. *Bioinformatics*, *28*(5), 628-635.

Statistics/Accuracy Demands for a Million Cancer Genome Warehouse

Standards for accuracy are a requirement for any assay that is to be used in treatment decisions and are a necessity for large-scale research as well. Current standards for DNA sequencing accuracy have been set primarily by the 1000 Genome Project and pertain only to the sequence of germline DNA. The international community is still in the process of establishing standards for the reporting of somatic mutations in cancer tissues relative to the patient's germline genome. The Cancer Genome Atlas in the US and the larger International Cancer Genome Consortium are leading this effort.

Setting standards is necessarily an ongoing task because the underlying technology by which the DNA is read is rapidly changing. Currently technology from three companies is creating the bulk of the data (Illumina, Complete Genomics, and Life Technologies), but newer companies with qualitatively different kinds of raw data such as Oxford Nanopore and Pacific Biosciences either have begun or will soon begin to produce data. On top of this, many cancer genome sequencing projects are in the midst of a shift from the current reliance on frozen tissue samples to tissue samples that are formaldehyde-fixed and paraffin-embedded (FFPE)[86]. This change is because in the U.S. in particular, the vast network of clinical practice is designed around FFPE storage. Essentially all U.S. archival samples are in this format. Handling and preservation of samples has a very significant effect on sequencing data that can be obtained from them, both for DNA and especially for RNA as it is considerably more labile; quantitative studies will be needed to precisely characterize these effects in the context of a very large database of sequenced samples from disparate sources.

Finally, new sequencing techniques that require many fewer cells than the millions required by current methods are undergoing rapid development[87]. These techniques are necessary for tissue samples obtained by fine needle biopsy, a safer and less invasive method of obtaining tumor tissue for analysis prior to major surgery and in cases where surgery is not appropriate. While it is feasible to obtain 100,000 cells from a fine needle biopsy[88], a fine needle biopsy often yields less than 100 cells by design, because the tumor is small, the physician wishes to specifically target a location within the tumor, or the physician wishes to target several

---

[86] Wagle, N., Berger, M. F., Davis, M. J., Blumenstiel, B., DeFelice, M., Pochanard, P., et al. (2012). High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discovery*, *2*(1), 82-93.

[87] Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, *472*(7341), 90-94.

[88] Glehr, M., Leithner, A., Gruber, G., Wretschitsch, P., Zacherl, M., Kroneis, T., et al. (2010). A New Fine-Needle Aspiration System. *Surgical innovation*, *17*(2), 136-141.

locations in separate biopsies in order to survey the heterogeneity within the tumor[89],[90]. Recovery of circulating tumor cells is another area where sequencing from very few cells is needed[91]. Accuracy standards for sequencing from a single cell are necessarily different from those for sequencing from many cells, because in preparing the DNA from a single cell, some significant percentage of it is inevitably lost, resulting in zero coverage over some regions of the reference genome. Thus, one cannot demand high or even moderate accuracy for all genes in the genome when sequencing from a single cell. If there is only one or very few cells per sequencing run, then data from several runs on different cells must be combined to get full coverage of the genome.

Use of a Reference Genome

Human germline genomes are about 99.9% identical and the places where there is high variability between people, called single nucleotide polymorphisms (SNPs), have nearly all been well-documented[92]. Still, there are about 10 million such positions in the reference genome. In terms of single nucleotide somatic changes, a patient's cancer genome (i.e. the dominant clone in the tumor) will differ from their germline genome in many fewer positions, somewhere between one[93] and a few tens of thousands[94], depending on the type of cancer. Thus when we read DNA either from the germline or the tumor, the majority of reads match a reference human genome, and at most a few times every million bases do we read something different (which is the reason why compression schemes are based on storing differences in the reads from reference). The problem of setting accuracy standards becomes when to decide that the something different that we read is worth noting because it might reflect a true difference, and when it is much too likely to be due to misread errors.

Why Uncertain Information Must be Tolerated and Stored

Even with millions of cells per sample, it is not sufficient to merely set a very high bar for the accuracy of cancer genome sequencing, e.g. tolerating only a few mistakes per genome. Standards must take into account the fundamental accuracy limitations in the technology and the extent to which different cells in the tumor tissue sample may have genetic differences that must be surveyed.

---

[89] Voit, C., Kron, M., Schäfer, G., Schoengen, A., Audring, H., Lukowsky, A., et al. (2006). Ultrasound-guided fine needle aspiration cytology prior to sentinel lymph node biopsy in melanoma patients. *Annals of surgical oncology*, *13*(12), 1682-1689.

[90] Micames, C. G., McCrory, D. C., Pavey, D. A., Jowell, P. S., & Gress, F. G. (2007). Endoscopic Ultrasound-Guided Fine-Needle Aspiration for Non-small Cell Lung Cancer Staging*. *Chest*, *131*(2), 539-548.

[91] Wicha, M. S., & Hayes, D. F. (2011). Circulating tumor cells: not all detected cells are bad and not all bad cells are detected. *Journal of Clinical Oncology*, *29*(12), 1508-1511.

[92] Altshuler, D. M., Lander, E. S., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., et al. (2010). A map of human genome variation from population scale sequencing. *Nature*, *467*(7319), 1061-1073.

[93] Lee, R. S., Stewart, C., Carter, S. L., Ambrogio, L., Cibulskis, K., Sougnez, C., et al. (2012). A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *The Journal of clinical investigation*, *122*(8), 2983.

[94] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, *446*(7132), 153-158.

Using current DNA sequencing technology, between 1% and 0.1% of the DNA bases will be misread on each read attempt, i.e., in each short read that is stored in a BAM file[95]. We will denote the fraction of time that a base is misread, the (single pass) *misread error rate*, as *p*. Newer technologies are exhibiting higher values of *p*, not lower, but compensate by providing increased speed, longer reads, easier DNA preparation, smaller input DNA requirement or some other advantage. So the range for *p* above is unlikely to change drastically. All methods make up for the inaccuracy in an individual read by reading DNA from the same position in the reference genome multiple times, to a particular depth of coverage *n*. But depth is never completely uniform across the genome, some regions may be duplicated but with only one or a few of the copies mutated, and many tumors harbor mutations that are only present is a small subset of the cells. The result is that many mutation calls made from DNA sequencing data are in a "grey zone" where we cannot be sure they are meaningful without further evidence, yet we cannot discard the data we have lest we lose our ability to recognize that further evidence when we get it. A quantitative analysis of error rates and their implications for mutation calling is given in Appendix B.

Reasoning with Uncertain Information

It is important to keep in mind that the purpose of building a genome warehouse is not merely to store and transport genomes, but to serve as the basis for a wide range of statistical inferences, from genomic inferences at the level of individual base pairs and structural polymorphisms, to genetic inferences regarding haplotypes and recombination events, to biochemical inferences regarding the pathways disrupted by various mutations, to medically-relevant inferences regarding correlations between genomic variation, phenotypes and treatments.

Uncertainty enters the process in the raw data provided by the sequencing platform, where quality scores attempt to quantify uncertainty in the output of the sequencing process, and and the uncertainty profile changes as more data is accumulated at a given position, as discussed above. It is essential to propagate uncertainty throughout subsequent stages of processing. For the crucial end-to-end inferences regarding genotype/phenotype correlations, there will rarely be enough data (enough cases of a specific disease with phenotypes) to be able to assess with certainty which aspects of the genotype are critical. That is, inference for cancer genomics will necessarily take place in a low signal/high noise regime, and it is essential to take this fact into account in designing data analysis pipelines.

Indeed, while it is often tempting to try to reduce uncertainty by making "calls" that convert uncertain quantities into discrete decisions, it is our belief that such calls should be postponed until as late as possible in pipelines, or forgone entirely.

Building a statistically-aware genome warehouse raises a number of challenges. First, there are computational challenges involved in computing and propagating error bars on uncertain quantities. There exist statistical tools based on resampling and Monte Carlo that can perform

---

[95] Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443-451.

such computations in principle, but such tools have rarely been applied at the scales that we envision. Moreover, there are numerous calibration issues that must be faced in converting quality scores and other genomic and genetic assessments of uncertainty into statistically meaningful error bars. Second, relational database operations in the genomic warehouse need to be statistically grounded. Specifically, while database joins are often optimized on the basis of raw operation counts, in a statistical setting the goal is that of reducing uncertainty by joining complementary pieces of data. Thus, optimizations should take into account the error bars of each individual source as well as joint measures of uncertainty. Finally, the ultimate goal of genome-based biology is to be able to establish causality, and thus the statistical underpinnings of causal inference (distinctions between experimental and observational data, attention to sampling frames, etc) need to be provided in the genome warehouse.

Potential Paths to a Million Cancer Genome Warehouse

We hope by this point you are convinced that it is technically feasible to build a Million Cancer Genome Warehouse, and that we could afford to build it in 2014. As it will take time to develop the right algorithms, APIs, and tools that can process this quantity of data efficiently and accurately, we argue that we need to get started soon rather than wait until the cost of sequencing drops, forcing us to innovate while trying to cope with a tsunami of sequencing data. Hence, to deliver on the vision of Precision Medicine, we must give researchers access to cancer genomic information and patient records at large scale as soon as possible.
Once the initial systems demonstrate benefits in understanding and treating cancer, we believe the technology will naturally become more widespread.

In order to get this process started, we will need a sufficient number of targeted panels, exomes/ genomes and patient records that researchers could access so as to require development of the technology we need.

The first candidates to have such data would be nations that already have electronic medical records (EMR) as part of national health care systems. There has already been discussion in Denmark, for example, about sequencing the DNA of its 6 million citizens. To estimate the number of new cancer patients each year and the number living with cancer, we extrapolate from the US statistics. About 0.5% (1.6M/320M in US) or 30,000 Danes may get cancer each year and about 4% (13M/320M) or 240,000 could be living with cancer. To get to a million genomes, we would need more countries in the EU to participate. China obviously has a much larger population, but has only recently started a national healthcare system, so it presumably would take several years to get to a million cancer genomes.

Within the US, there is no question that the fastest progress could be made if all the data was in a single large scale repository that gave researchers unfettered access. However, this option may not be politically feasible in the US until the benefits of the warehouse are demonstrated. If so, how could we make demonstrate the value of a million cancer genome warehouse? Quoting Shimon Peres:

If a problem has no solution, it may not be a problem, but a fact - not to be solved, but to be coped with over time.

We see three paths to cope with the difficulty of creating a single US repository:

1. *Top Down*. A US government agency to start its own trial with, say, 1000 patients, and use the trial to drive a standard for patient records. Given the time for standardization and to run the study, it would take perhaps five years to get to a thousand cancer genome warehouse, and then another several years extend and modify these efforts to scale to a million genomes. This seems like a sensible, conservative plan forward, with a potentially large warehouse of clean data in 2020 or later.
2. *Patient Push*. Create a public resource, provide a common consent form, and appeal to patients to donate their cancer data, perhaps as part of a non-profit organization. This option would be to appeal to patients directly and would create a prominent role for patient advocacy groups insisting that their medical data be released by the institutions treating those who participate. Advocacy groups might even help transcribe electronic medical records (EMR) from the many centers into a standard form that this organization could use. Several consent forms geared toward this approach have been developed, most prominently the Portable Consent developed by John Wilbanks and used by Sage Bionetworks.
3. *Center Push*. Instead of dealing directly with cancer patients, another approach is to work with centers treating cancer patients. The idea would be to form a consortium starting with a small number of enlightened centers that 1) already have an EMR, 2) already plan to do full genome sequencing, and 3) are interested in sharing data in a large repository so as to improve the care of their patients. For example, MD Anderson treats ~110,000 patients per year, with a third being new, or about 40,000 new patients per year. That represents about 2.5% of all new cancer patients in the US. The second largest is Memorial Sloan–Kettering Cancer Center, which is similar but slightly smaller. While there are 1500 cancer centers in the US, the few largest centers combined treat 10% to 20% of all US patients, or 160,000 to 320,000 new patients and 1.3M to 2.6M people living with cancer. This option presumes that such centers would be willing to cooperate to help their patients as well as to make progress on fighting cancer. Investigators combine data when forced to to gain necessary statistical power. The experience in genome-wide association studies is illustrative. For example, initial studies searching for statistically significant genetic links to schizophrenia did not have enough patients to find a link. The NIH helped inspire cooperation between centers in the US to find the link, but this effort was also short of data. It was only when they formed the International Schizophrenia Association--involving 106 researchers at 26 institutions in 8 countries--that they pooled enough information to make a significant discovery.[96] As mentioned above, centers would need to have ways to regulate information usage until the studies have been completed and the papers published.

[96] Stefansson, H. et al. Common variants conferring risk of schizophrenia. *Nature*. 2009 August 6; 460(7256): 744–747.

These last two options would not have the pristine structured data of the first option. However, if the first option is not going to produce a database until after 2020, then the Google authors of "The Unreasonable Effectiveness of Data" would likely expect these paths to be at least as successful as the first one as they will generate a lot more data sooner. From our perspective, we can see the benefits or proceeding on all three fronts, rather than betting on a single approach.

Note that if warehouses deliver on their potential, a new industry would likely quickly grow around the processing and analysis of cancer genetic information. Being early is often a big advantage in a new industry, and whatever country becomes the center of that industry will likely benefit from new jobs and taxes.

Conclusion

This white paper makes the argument that a Million Cancer Genome Warehouse is both important for society as a major step towards achieving Precision Medicine and that advances in technology have made it feasible. While we need to improve the cost-performance and accuracy of software that makes the warehouse useful and address policy issues around data access permissions and interfaces, we believe society can now easily afford to build such a warehouse. To develop new algorithms and software to make inroads in the fight against cancer using genetic information, we need access to such a repository soon before we are swamped with the demands for processing and storing millions of genomes that will surely appear as the price of sequencing drops.

In countries like the US without a national healthcare system, it will be much more difficult to create a single repository until we have definitive evidence of its benefits. We recommend pursuing three paths: a large government clinical trials driven approach, a patient-oriented organization, and a medical center-oriented consortium. While medical institutions are notoriously slow to move, we believe there are several points which will be important in facilitating this process:

- Common consent form for all patients that contribute to the database in order for collection and dissemination of standard data from different treatment centers.
- Aligning the interests of clinicians and cancer researchers who will use the database in order to generate interest and motivate institutional adoption.
- Provide an early example of the utility of this database, most likely at a single medical center, to promote adoption among additional centers.
- Mobilizing patient advocacy groups in order to promote adoption among medical centers and to ensure the proper transfer of patient information.

We urge the extension of existing and the piloting of new large medical genomic database efforts around the world to investigate their potential. If they prove to help researchers discover new ways to combat cancer or other diseases and help medical practitioners to save the lives of our fellow man, which we expect, some will surely migrate to a self-sustaining model.

Appendix A: The Universal Human Reference and Representation of Genome Structural Variation

Currently the human reference genome (GRCh37) is largely composed of what is termed a “golden path” set of sequences, representing, barring a few gaps, a complete monoploid set of chromosomes. Its importance to biomedical research is hard to overstate, ultimately providing a proxy to a universal coordinate system for human genetics. However, the golden path representation has a number of important drawbacks as a reference[97].

Many subsequences prevalent in the population have either been lost (deleted) from the genomes used to build the golden path or gained in other human genomes (inserted). Such subsequences are described as insertions with respect to the golden path, and while any such insertion can be described by its insertion point and sequence, the golden path coordinate system does not readily extend to describing nested variations contained within these insertions. For example, and most obviously, single nucleotide polymorphisms (SNPs) themselves present within such insertions have no coordinate within the golden path. Additionally, the current golden path, representing a chimera of haplotypes for each chromosome, has numerous minor alleles that are rare in the population, both in terms of SNPs and structural variations. This is true for both BAM and VCF files, where unmapped reads are a storage headache and ignoring the variants they contain leads to problematic biases. Finally, many tools use the golden path sequences (reference mappers[98],[99] , primer software[100], reference based assemblers[101],[102], peak callers[103] for chromatin immunoprecipitation assays that employ sequencing, etc.), and much annotation effort has been dedicated to the golden path, creating a disturbing bias in research towards reference allele discovery and curation at the expense of the alleles not represented in the reference[104]. As the reference is currently derived largely from a European individual[105], the reference allele bias is also a potential

---

[97] Mardis, E. R. (2010). The $1,000 genome, the $100,000 analysis. *Genome Med*, *2*(11), 84.

[98] Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

[99] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, *10*(3), R25.

[100] Tusnady, G. E., Simon, I., Varadi, A., & Aranyi, T. (2005). BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic acids research*, *33*(1), e9-e9.

[101] Pop, M., Phillippy, A., Delcher, A. L., & Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in bioinformatics*, *5*(3), 237-248.

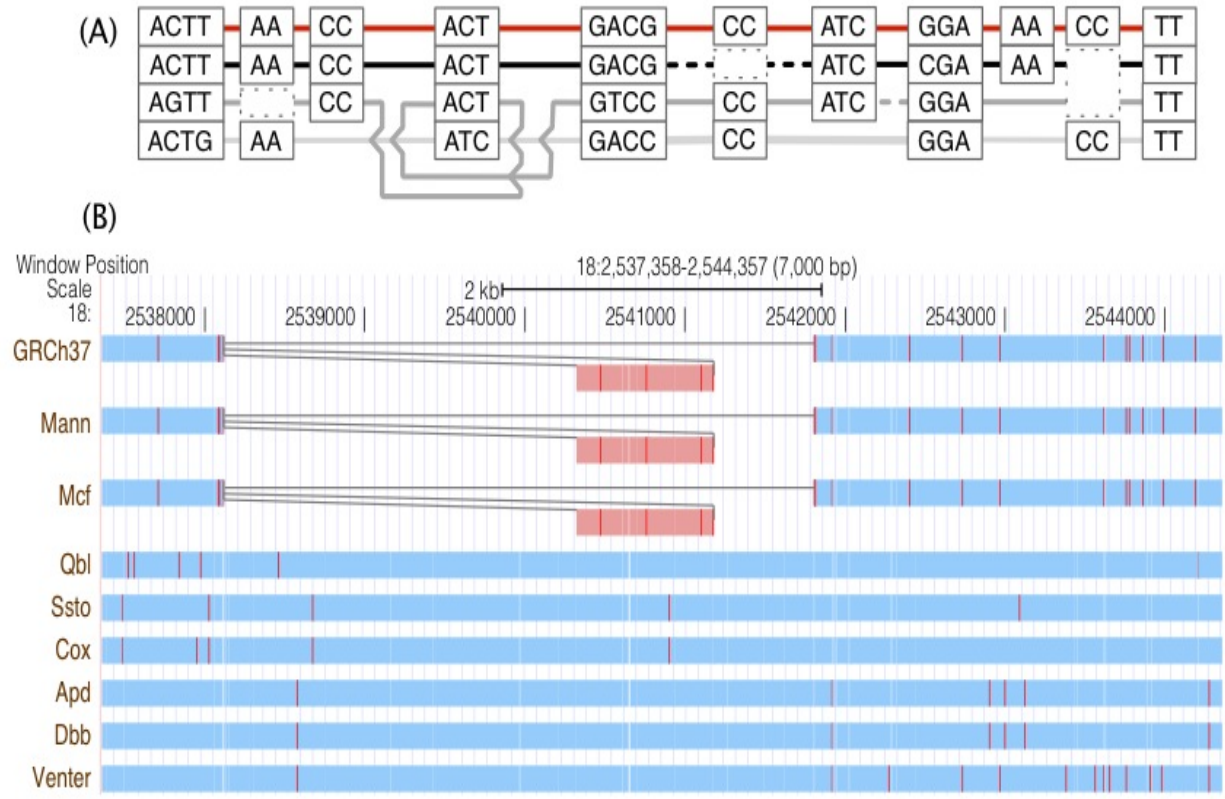
[102] Gnerre, S., Lander, E. S., Lindblad-Toh, K., & Jaffe, D. B. (2009). Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome biology*, *10*(8), R88.

[103] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, *9*(9), R137.

[104] Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., et al. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol*, *7*(Suppl 1), S4.

[105] Osoegawa, K., Mammoser, A. G., Wu, C., Frengen, E., Zeng, C., Catanese, J. J., et al. (2001). A bacterial artificial chromosome library for sequencing the complete human genome. *Genome research*, *11*(3), 483-496.

ethnic bias, which may cause greater problems when working with individuals from other sub-populations.



(A) A schematic diagram illustrating the fat consensus reference as a path through an alignment of haplotype sequences. Stacked boxes represent aligned, oriented subsequences. Lines represent connections between subsequences, rows of alternating boxes and lines represent contiguous sequences. In this toy example three haplotypes are shown in black, grey and light grey. The dotted lines represent positions of uncertainty (gaps) within the sequences, as is typically present within assemblies or when a genome is composed of a large number of variants interspersed with gaps. The red sequence represents the fat consensus reference. It includes a consensus representation of every possible aligned subsequence, thus every position is given a coordinate within the reference. (B) A prototype browser display in which haplotypes are aligned to the coordinate system of a fat reference. The blue/red bars represent sequence. The red ticks show substitutions with respect to the consensus, the lines represent connections as in (A). The example shows a complex inversion/deletion that spans about 5 kilo bases. The GRCh37 sample, the current golden path, carries the minority allele for this rearrangement, while the fat reference has the majority allele.

To address these issues, the Genome Reference Consortium[106] is identifying and sequencing alternative haplotypes in highly polymorphic regions and including these along with the reference genome. For example, further haplotype sequences for the Major Histocompatibility Complex (MHC)[107] have been added to the reference (though still only one is designated as being on the golden path). This approach necessitates a homology map between the multiple

[106] Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., et al. (2011). Modernizing reference genome assemblies. *PLoS biology*, *9*(7), e1001091.

[107] Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R. J., Almeida, J., et al. (2008). Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics*, *60*(1), 1-18.

possible references to avoid chronic ambiguities in the multiple mapping of both reads and annotations between functionally equivalent and evolutionary homologous regions. It creates a protocol that is inconsistent with the simple linear coordinate system currently used by a wealth of existing genome analysis tools, including all those that support the BAM and VCF formats.

To resolve this, we propose creating a graph structure called a *sequence graph* based on the common variations of the reference genome as identified by the Genome Reference Consortium and collaborators, and then to produce a canonical linearization of this graph structure informally called a "fat consensus" golden path[108]. The fat consensus traverses the graph of all variation, defining a linear set of chromosomes (see figure below). We call it fat because we propose it should be inclusive, so that its coordinates provide space for all suitably prevalent insertions. It is a consensus, in that it should describe as far as possible the most common path of variation present in the population, to make it as representative as possible. The coordinate system established by the fat consensus golden path could be used at all levels of the system, from the short read mappings in the BAM files to the highest level relational database.

The sequence graph contains more information than the fat consensus, and will itself be useful for many studies[109]. Genome variants are frequently inherited together, and often in large blocks that may, as a haplogroup (common haplotype shared by individuals), represent a large region that is identical by descent in cells within an individual, in individuals of a family or even in a whole sub-population of individuals. Because it represents the fundamental structure of genetic variation, this information on haplogroups is important to research and medicine. With newer sequencing technologies that produce longer reads, and the development of clever barcoding methods to better organize short reads, reliable haplotype information will be increasingly available, and will at some point become the norm[110],[111],[112]. A million genome database for storing the next generation of DNA sequence data should be designed to take haplotype information into account.

By giving haplogroups a place in the data structure, they can themselves be tiled together to describe, in a form of genetic shorthand, much of an individual genome. The general graph-theoretic representation of a sequence graph also allows the compact and computable representation (as subgraphs) of complex structural rearrangement events that occur in

---

[108] Paten, B., Nguyen, N., Zerbino, D., Earl, D., Raney, B., Hickey, G., Diekhans, M., Haussler, D. The Reference Problem: Defining a Consensus DNA Sequence from a Population. In preparation.

[109] Paten, B., Zerbino, D., Haussler, D., Genome Histories. In preparation.

[110] Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., & Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences*, *106*(19), 7702.

[111] Fan, H. C., Wang, J., Potanina, A., & Quake, S. R. (2010). Whole-genome molecular haplotyping of single cells. *Nature biotechnology*, *29*(1), 51-57.

[112] Roach, Jared C et al. "Analysis of genetic inheritance in a family quartet by whole-genome sequencing." *Science* 328.5978 (2010): 636-639.

cancer, such as chromothripsis[113], with associated uncertainty. This information is expressed in the VCF format file as raw observations of unexpected joins and atypical read depth in DNA from that tumor that maps to particular locations in the reference genome. A sequence graph interprets this information from the VCF file as a graph of tumor structural variation. Classification of rearrangements within an ontology of types such as inversions, translocations, etc., can be built by recognising characteristic subgraphs. This could facilitate the study of gene fusions, amplifications and losses, all fundamental to cancer, putting methods in this area on a par with the more well-developed methods for single nucleotide variants. Whatever precise format is chosen to represent structural variation, it will be important that the million genome database not treat structural changes as second-class citizens in the universe of cancer genome mutations.

[113] Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, *144*(1), 27-40.

Appendix B: Expected Number of False Positives Somatic Mutation Calls Per Position in the Reference Genome Due to Sequencer Read Error, and Its Implications

Let us assume for simplicity that we know the germline genome of the patient perfectly, treating it as the reference, and are sequencing from their cancer biopsy. We cannot avoid a substantial number of misread errors when reading DNA from the tumor that will give the appearance of mutations when there are none. Even at a misread error rate of p = 0.0001 (i.e. 0.01% of the bases in the read are wrong, 10 times lower than the average that can be expected from typical technology), there will be roughly 300,000 single base misreads (*false positives*) already with n = 1 coverage of a genome with 3 billion bases, and proportionally more for higher coverage (See Error Characteristics Table below).

Read Depth per Patient n	Error rate p	Probability of at least 1 misread at given site $(1-(1-p)^n)$	Expected number of sites out of 3 billion with at least 1 misread	Required redundancy for a single genome of 3 billion sites x and (x/n)	Required redundancy for a cohort of 1,000 patients, 3 billion sites x and $(x/1{,}000n)$
500	0.0001	0.0488	146,000,000	5 reads (0.01)	49 reads (0.000098)
500	0.001	0.394	1,180,000,000	8 reads (0.016)	256 reads (0.000512)
500	0.01	0.993	2,980,000,000	16 reads (0.032)	1,933 reads (0.00387)
100	0.0001	0.00995	29,900,000	4 reads (0.04)	21 reads (0.00021)
100	0.001	0.0952	286,000,000	6 reads (0.06)	77 reads (0.00077)
100	0.01	0.634	1,900,000,000	9 reads (0.09)	456 reads (0.00456)
30	0.0001	0.00300	8,990,000	4 reads (0.133)	13 reads (0.000433)
30	0.001	0.0296	88,700,000	5 reads (0.167)	37 reads (0.001233)
30	0.01	0.260	781,000,000	7 reads (0.233)	171 reads (0.0057)

Error Characteristics Table. DNA sequencing error characteristics at different read depths as discussed in the text.

Assuming that errors are uniform, independent, and occur at rate *p*, we can determine the probability that there will be various numbers of misreads that falsely produce non-reference bases that are not actually present in the DNA sequence being read. Let *Q(x,n,p)* be the probability that there are at least *x* misreads falsely producing the same non-reference base at a particular site when the sequencing depth is *n* and the single pass misread error rate is *p*. We can upper bound this value by

$$Q(x,n,p) \leq 3 \sum_{k=x}^{n} \binom{n}{k} (1-\frac{p}{3})^{(n-k)} (\frac{p}{3})^{k}$$

This bound is fairly tight when the number being bounded is small (see "Upper bound approximation" table below for probabilities of false positives for values of the upper bound approximation and the actual value of the probability). Let *M(x,n,p)* = *Q(x,n,p)* * 3,000,000,000, i.e. *M(x,n,p)* is the expected number of false positives in which *x* or more misreads falsely produce the same non-reference base at some site when sequencing a whole genome. If *x* is large enough so that *M(x,n,p)* < 1 then we can strongly trust a mutation call based on *x* reads of the same non-reference base at a given position because we don't expect this to happen by chance even once in the entire genome of 3 billion bases. We call the smallest such *x* the *required redundancy*. This is a very high bar for accuracy; more false positives than this can usually be tolerated in practice. Still, it serves as a useful benchmark.

The Error Characteristics Table above gives some values of the required redundancy *x* for several cases of *n* and *p* (Column 5). In parentheses we indicate what fraction of the total coverage at a particular position is in the required redundancy. For example in the first row we see that if the error rate is *p* = 0.0001 and the patient's DNA is sequenced to a depth of 500 at a given position then the required redundancy is 5 reads (i.e. 5 out of 500 reads, or 1%). Five is the number of times we would need to see a given non-reference base in that position in order to have a reliable mutation call.

If we have a cohort of patients, and we are interested to know if the same mutation is either very dominant in one or a few patients and/or low level but recurrent among many patients in a given position of the genome, then we can add up the counts from all reads in all patients in a given position. This leads us to considerably smaller required redundancies as a percentage of the total number of reads in all patients at the given position. These required redundancies are given in Column 6 of the Error Characteristics Table for a cohort of 1,000 patients.

Note that the above analysis only addresses the issue of whether we can be confident that the observed differences between the reference and the tumor are real (due to actual somatic mutations) or if they may be due to misreads. Even if the differences between the patient's tumor DNA and germline DNA are real, they may still only reflect random "passenger" mutations in the tumor that have no bearing on the disease. In a large cohort, if the same mutation occurs in many different patients, we can be more confident that it is a driver rather than a passenger. However, some regions of the genome are more prone to random mutation, and this must be

taken into account when making higher level causal inferences of this type[114].

The values of the required redundancy reveal our current inability to detect mutations that occur in a small fraction of the cells in the tissue sample in a single patient. At n = 30-fold coverage, a mutation that is homozygously present in 5% of the cells (or heterozygously present in 10% of the cells in the tissue sample) has only a 6% chance of having the required redundancy to be reliably detected even in the most accurate case of misread rate p = 0.0001. A mutation homozygously present in 1% of the cells has only a 0.02% chance of having the required redundancy. For n = 100-fold coverage, the corresponding values are 74.2% and 1.8% respectively. It is not until we reach n = 950-fold coverage, where 6 reads are required to avoid false positives, that a mutation homozygously present in 1% of the cells (heterozygous in 2% of the cells) has more than a 90% chance of having the required redundancy. Thus, unless we are prepared to accept quite a number of false positives or to do a second phase of additional validation on the calls we make, as is done in the TCGA project, it is not possible to reliably call mutations that only occur is a small fraction of the cells of a single genome. It does become possible at the cohort level, however. To do this, we have to store and reason with isolated read discordancies in a patient, at frequencies far below the required redundancy, even though in any one tumor these may be overwhelmingly likely to be misread errors.

The above analysis is highly simplified in that it does not address structural variation in the cancer genome (see Appendix A). Not only is structural variation difficult to precisely capture and assess, but the presence of duplications of genomic regions, with some copies mutated and other copies not mutated, makes assessing single nucleotide changes in those regions much more difficult.

There are also technical errors in base calling that are not independent and can wreak havoc. These are often due to systematic mismapping of the reads to the reference genome. When the same mapping error is repeated multiple times, we get several reads that exhibit the false appearance of independent evidence for a particular difference relative to the reference genome. Both the 1000 Genomes Project and the TCGA/ICGC projects have found this to be a major source of errors among the variant calls that otherwise appear to be made with high confidence. In a database of 10^{17} data points (1 million genomes, each with 3 billion sites and with 30-40 single pass base calls at each site), any systematic, non-independent errors of this type must be uncovered and either eliminated or incorporated into the model of uncertainty.

[114] (2011). The Spectra of Somatic Mutations Across Many Tumor Types. Retrieved July 22, 2012, from http://www.genome.gov/Multimedia/Slides/TCGA1/TCGA1_Lawrence.pdf.

Upper bound approximation table. The table below gives an estimated upper bound approximation (formula in the main text) and actual probabilities of a false positive at a single position for various coverages and error rates when $x = 1, \ldots, 16$ identical non-reference reads are required. At low numbers, the estimation very closely bounds actual numbers.

n	*p*	est. *x* = 1	actual *x* = 1	est. *x* = 2	actual *x* = 2	est. *x* = 3	actual *x* = 3	est. *x* = 4	actual *x* = 4
500	0.0001	0.050	0.049	4.1e-4	4.1e-4	2.3e-6	2.3e-6	9.4e-9	9.4e-9
500	0.001	0.46	0.39	0.037	0.037	0.0020	0.0020	8.4e-5	8.4e-5
500	0.01	1	0.99	1	0.97	0.70	0.55	0.26	0.24
100	0.0001	0.010	0.0099	1.6e-5	1.6e-5	1.8e-8	1.8e-8	1.4e-11	1.4e-11
100	0.001	0.098	0.095	0.0016	0.0016	1.8e-5	1.8e-5	1.4e-7	1.4e-7
100	0.01	0.85	0.63	0.13	0.13	0.014	0.014	0.0011	0.0011
30	0.0001	0.0030	0.0030	1.4e-6	1.4e-6	4.5e-10	4.5e-10	0	0
30	0.001	0.030	0.030	1.4e-4	1.4e-6	4.5e-7	4.5e-7	1.0e-9	1.0e-9
30	0.01	0.29	0.26	0.014	0.014	4.2e-4	4.2e-4	9.5e-6	9.5e-6

n	*p*	est. *x* = 5	actual *x* = 5	est. *x* = 6	actual *x* = 6	est. *x* = 7	actual *x* = 7	est. *x* = 8	actual *x* = 8
500	0.0001	3.1e-11	3.1e-11	0	0	0	0	0	0
500	0.001	2.7e-6	2.7e-6	1.8e-9	1.8e-9	3.6e-11	3.6e-11	0	0
500	0.01	0.082	0.080	0.022	0.021	0.0049	0.0049	9.9e-4	9.9e-4
100	0.0001	0	0	0	0	0	0	0	0
100	0.001	9.1e-10	9.1e-10	4.8e-12	4.7e-12	0	0	0	0
100	0.01	7.1e-5	7.1e-5	3.8e-6	3.8e-6	1.7e-7	1.7e-7	6.5e-9	6.5e-9
30	0.0001	0	0	0	0	0	0	0	0
30	0.001	1.7e-12	1.7e-12	0	0	0	0	0	0
30	0.01	1.6e-7	1.6e-7	2.6e-11	2.6e-11	0	0	0	0

n	*p*	est. *x* = 9	actual *x* = 9	est. *x* = 10	actual *x* = 10	est. *x* = 11	actual *x* = 11	est. *x* = 12	actual *x* = 12
500	0.0001	0	0	0	0	0	0	0	0
500	0.001	0	0	0	0	0	0	0	0
500	0.01	1.8e-4	1.8e-4	2.9e-5	2.9e-5	4.2e-6	4.2e-6	5.6e-7	5.6e-7
100	0.0001	0	0	0	0	0	0	0	0
100	0.001	0	0	0	0	0	0	0	0
100	0.01	2.2e-10	2.2e-10	6.7e-12	6.7e-12	0	0	0	0
30	0.0001	0	0	0	0	0	0	0	0
30	0.001	0	0	0	0	0	0	0	0
30	0.01	0	0	0	0	0	0	0	0

n	*p*	est. *x* = 13	actual *x* = 13	est. *x* = 14	actual *x* = 14	est. *x* = 15	actual *x* = 15	est. *x* = 16	actual *x* = 16
500	0.0001	0	0	0	0	0	0	0	0
500	0.001	0	0	0	0	0	0	0	0
500	0.01	7.0e-8	7.0e-8	8.1e-9	8.1e-9	8.7e-10	8.7e-10	8.8e-11	8.8e-11
100	0.0001	0	0	0	0	0	0	0	0
100	0.001	0	0	0	0	0	0	0	0
100	0.01	0	0	0	0	0	0	0	0
30	0.0001	0	0	0	0	0	0	0	0
30	0.001	0	0	0	0	0	0	0	0
30	0.01	0	0	0	0	0	0	0	0

Appendix C: Alternate Design of Warehouse Hardware

AIn this section, we assume an alternate design, in which the storage is distributed across all servers. Similarly, we consider an OCP rack design, where each rack hosts 18 2U servers. We assume each server contains 2 processors, 512 GB DRAM, and 12 disks of 5TB. Using same prices as before (i.e., $5K/TB DRAM, $30/TB Disk, $200 per processor), we get $5K per server. Adding the price for motherboard and rounding it up we get $6K per server. Thus, each rack has 360 cores, 1.08PB, and 9TB of DRAM at $108K. To store 125 PB, we need 116 racks, at a cost of $12.5M. In addition to 125 PB of disk storage, the 116 racks host over 40,000 cores, and 1 PB of DRAM.

Typically, the cost of routers and switches is around 25% of server cost, or about $3M. In addition, we assume the cabling cost is $2K/rack or $332K for all racks.

The CAPEX, which includes servers, networking equipment, cabinets and cabling, is $16M. If we add 10% to account for various estimates and larger number of racks, which will have higher hosting costs, we come close to the $18M of the earlier estimate.

Appendix D. Technical Approaches to Germline DNA Privacy Concerns

In addition to somatic DNA changes that occur only in the tumor cells, inherited DNA variations in all of the patient's cells can affect how an individual develops a particular cancer and how the disease responds to treatment regimes, so it will also be important to share this information while protecting patient privacy.

An important goal for a Million Cancer Genome Warehouse is performing genome-wide association studies (GWAS) of both common and rare inherited single nucleotide polymorphisms (SNPs) to compare regions of the genome between matched cohorts with and without a particular cancer. The large scale of the warehouse will increase the ability to achieve desired levels of statistical power for detecting weak associations between single nucleotide or structural variants and cancer susceptibility while limiting the rate of false positives. Achieving this goal is a critical step in realizing the vision of personalized medicine.

However, for the germline mutations, as discussed above, patient privacy is a major concern and the warehouse's large scale introduces a major design challenge. Previously, many studies have pooled individual genotype data together while making the allele frequencies of each SNP in the pool publicly available. It has been implicitly assumed that releasing such summary data provides a secure way to share the results of a study without compromising the privacy of the study participants. However, Homer et al.[115] showed that it is possible, by examining datasets based on high-density SNP arrays, to accurately detect the presence of individual genotypes in a mixture of pooled DNA even when each individual's DNA is present in only small concentrations. Although this analysis was aimed at applications in forensic science, these findings raised the possibility that the presence of individual genotypes could be inferred from summary data, and this possibility has led to the removal of publicly available summary data from previous studies as a conservative means of protecting the privacy of human subjects[116].

A key question about the types of cancer genomic analyses that researchers will want to perform is whether it will be sufficient to have access to summary SNP data for only a subset of the SNPs ("exposed" SNPs). In cases where such access is sufficient, then it is likely that an appropriately defined level of privacy can be maintained if the number of exposed SNPs is sufficiently small. Establishing privacy guidelines of this kind requires an understanding of how the number of exposed SNPs varies as a function of factors such as the allele frequencies of the SNPs, the number of individuals in the DNA pool and, of particular importance, the method used to detect the individual in the pool. An analysis of this kind was pursued by Homer et al[117], who proposed a particular detection method and estimated the statistical power of detecting

---

[115] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, *4*(8), e1000167.

[116] Gilbert, N. *Nature* doi:10.1038/news.2008.1083 (4 September 2008).

[117] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, *4*(8), e1000167.

an individual genotype in a sample of exposed SNPs using this new method. But although an analysis of any specific detection method provides an estimate of its power of detection, it remains possible that another method could provide increased power and that therefore no guarantee could be provided that its power of detection was below some acceptable level.

Sankaraman et al[118] propose a method and a tool, SecureGenome, that provides an upper bound on the detection power achievable by any method, which yields guidelines as to which set of SNPs can be safely exposed for a given pool size with a maximal allowable power Beta and false-positive level Alpha. SecureGenome is based on a statistical hypothesis testing formulation, for which the Likelihood Ratio test (LR test) attains the maximal power achievable. This approach provides a guarantee that it will be safe to expose a set of SNPs for which the LR test does not achieve sufficient power. SecureGenome takes as input a genotype dataset (including the individuals' genotypes), a reference dataset and a ranking of the SNPs, removes SNPs that are in linkage disequilibrium, and determines the number of highly ranked SNPs that can be safely exposed. The program outputs this value along with the power of the LR test evaluated both empirically and theoretically.

In the context of a Million Cancer Genome Warehouse, the approach used by this tool may be useful in developing a methodology that enables effective cancer genomics research while taking into account both privacy and the need to leverage data collected throughout the community.

[118] Sankararaman, S., Obozinski, G., Jordan, M. I., & Halperin, E. (2009). Genomic privacy and limits of individual detection in a pool. *Nature genetics*, *41*(9), 965-967.

Author Bios

David Haussler (UC Santa Cruz) is currently Director of the Center for Biomolecular Science and Engineering at UCSC, co-Director of the California Institute for Quantitative Biosciences representing UCSC, and adjunct professor at Stanford and UCSF. He develops new statistical and algorithmic methods to explore the molecular function, evolution, and disease process in the human genome, integrating comparative and high-throughput genomics data to study gene structure, function, and regulation. He is credited with pioneering the use in genomics of hidden Markov models, stochastic context-free grammars, and discriminative kernel methods, the latter first applied to gene expression in cancer in 1999. As a collaborator on the international Human Genome Project, his team posted the first assembly of the human genome sequence on the Internet on July 7, 2000. They subsequently developed the UCSC Genome Browser, a web-based tool that is used by hundreds of thousands of biomedical researchers worldwide in both basic and translational research, receiving more than 3 million page requests per week. Most recently, he built the CGHub database to hold the National Cancer Institute's cancer genome data. His group's informatics work on cancer genomics provides a complete analysis pipeline from raw DNA reads through the detection and interpretation of mutations and altered gene expression in tumor samples. He works with the Stand Up To Cancer "Dream Teams" of the American Association for Cancer Research and The Cancer Genome Altas project of the NCI to discover molecular causes of cancer and pioneer a new personalized, genomics-based approach to cancer treatment. Elected to the National Academy of Sciences and the American Academy of Arts and Sciences in 2006, he has received the 2011 Weldon Memorial prize for application of mathematics and statistics to biology, 2009 ASHG Curt Stern Award in Human Genetics, the 2008 Senior Scientist Accomplishment Award from the International Society for Computational Biology, the 2006 Dickson Prize for Science from Carnegie Mellon University, and the 2003 ACM/AAAI Allen Newell Award in Artificial Intelligence.

David Patterson (UC Berkeley) is the Pardee Professor of Computer Science, Director of the Parallel Computing Laboratory (Par Lab), and a founding member of the Algorithms, Machines, and People Laboratory (AMP Lab). In the past, he served as Director of the Reliable and Distributed Systems Laboratory (RAD Lab), as Chair of Berkeley's CS Division, Chair of the Computing Research Association, and President of the Association for Computing Machinery (ACM), which at 108,000 members is the largest professional society in information technology. His most successful projects have been Reduced Instruction Set Computers (RISC), Redundant Arrays of Inexpensive Disks (RAID), and Network of Workstations (NOW), which a recent NAE report cited as helping lead to three multibillion dollar industries. This research led to many papers, six books, and more than 35 honors, some shared with friends, including election to the National Academy of Engineering, the National Academy of Sciences, and the Silicon Valley Engineering Hall of Fame. He was named Fellow of the Computer History Museum and both AAAS organizations. From the University of California he won the Outstanding Academic Alumnus Award (UCLA College of Engineering) and the Distinguished Teaching Award (UC Berkeley). From the ACM, where as a fellow, he received the Distinguished Service Award, the

Karlstrom Outstanding Educator Award, the SIGARCH Eckert Mauchly Award, the SIGMOD Test of Time Award, and the SIGOPS Hall of Fame Award. He is also a fellow at the IEEE, where he received the Johnson Information Storage Award, the Undergraduate Teaching Award, and the Mulligan Education Medal. He shared the IEEE von Neumann Medal and the NEC C&C Prize with John Hennessy, President of Stanford University and co-author of two of his books. His most recent award is the 2012 Jean-Claude Laprie Award in Dependable Computing from IFIP.

Mark Diekhans (UC Santa Cruz) is Technical Director of the CGHub repository for NCI cancer genomics data and a bioinformatics software architect and engineer with the Center for Biomolecular Science and Engineering at UCSC. He has a wide range of experience building large, high-performance software systems, including a highly successful commercial operating system. His bioinformatics work includes human genome annotation, comparative genomics, and cancer genomics.

Armando Fox (UC Berkeley) is a Professor-In-Residence and a co-founder of the Berkeley RAD Lab (Reliable Adaptive Distributed Systems) and AMP Lab (Algorithms, Machines and People). His 2003 collaboration with David Patterson on Recovery-Oriented Computing earned him the distinction of being included in the "Scientific American 50" top researchers and led to the formation of the RAD Lab, where he was a coauthor of the influential position paper "Above the Clouds: A Berkeley View of Cloud Computing." This paper was one of the first widely-read academic articles that outlined a cloud computing research agenda, and led to several successful collaborations with top machine learning researchers on the application of machine learning to operational problems in cloud computing datacenters. Prof. Fox reinvented the undergraduate Software Engineering course at Berkeley and wrote a new textbook *Engineering Long-Lasting Software: An Agile Approach Using SaaS & Cloud Computing*, co-authored with David Patterson. Prior to joining Berkeley, he was an Assistant Professor of Computer Science at Stanford, where he received the National Science Foundation CAREER Award, the Robert Noyce Family Faculty Fellowship, the IBM Young Faculty Fellowship, and teaching awards from Stanford University (his former employer), the Society of Women Engineers, and Tau Beta Pi Engineering Honor Society. In previous lives he helped design the Intel Pentium Pro microprocessor and founded a small company to commercialize his UC Berkeley dissertation research on mobile computing. He holds degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, the University of Illinois at Urbana-Champaign, and UC Berkeley.

Michael Franklin (UC Berkeley) is the Thomas M. Siebel Professor of Computer Science, focusing on new approaches for large-scale data management and data analysis. At Berkeley he is Director of the Algorithms, Machines and People Laboratory (AMP Lab), a cross-disciplinary collaboration integrating machine learning, large-scale cluster computing and crowdsourcing to develop new approaches to big data analytics. He is also the Principal Investigator of a DARPA seedling project (a one year project awarded as part of the CRASH program), which supported foundational work that led to the establishment of the AMPLab. His on-going research projects are in the areas of data stream processing and continuous analytics,

scalable query processing, large-scale sensing environments, data integration, hybrid human/ computer data processing systems, and cross-data center consistency protocols. He is a Fellow of the Association for Computing Machinery and recipient of the NSF CAREER award, ACM SIGMOD “Test of Time” award, and Outstanding Advisor Award from the Computer Science Graduate Student Association at Berkeley. He is currently serving as a committee member on the U.S. National Academy of Sciences study on Analysis of Massive Data. He was the founder and CTO of Truviso, Inc. a real-time data analytics company that enables customers to quickly make sense of diverse, high-speed, continuous streams of information. Truviso was recently acquired by Cisco Systems to provide real-time analytics for Network Analytics and other data-in-motion applications. He received his Ph.D. in Computer Science from the University of Wisconsin, Madison in 1993.

Michael I. Jordan (UC Berkeley) is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. His research in recent years has focused on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in statistical genetics, signal processing, computational biology, information retrieval and natural language processing. Prof. Jordan is a member of the National Academy of Sciences, a member of the National Academy of Engineering and a member of the American Academy of Arts and Sciences. He is a Fellow of the American Association for the Advancement of Science. He has been named a Neyman Lecturer and a Medallion Lecturer by the Institute of Mathematical Statistics. He is an Elected Member of the International Institute of Statistics. He is a Fellow of the AAAI, ACM, ASA, CSS, IMS, IEEE and SIAM. He received the 2008 SIAM Activity Group on Optimization Prize and the 2009 ACM/AAAI Allen Newell Award in Artificial Intelligence.

Anthony D. Joseph (UC Berkeley) is an Associate Professor and holds a UC Berkeley Chancellor's Professorship. He received his Ph.D. degree in computer science from MIT in 1998, and is a member of IEEE, ACM, and USENIX. He was Director of Intel Labs Berkeley, Intel Corporation's research laboratory in Berkeley from 2008 to 2011, a Massachusetts Institute of Technology Martin Luther King Jr. Visiting Scholar in 2004, a Nokia Foundation Visiting Fellow, in 2004, a member of the DARPA-sponsored Defense Science Study Group from 2004 to 2005, awarded a National Science Foundation CAREER Award in 2000, received an Okawa Foundation Research Grant in Telecommunications and Information Processing in 1999, and received an IBM Faculty Development Award in 1999. He is developing adaptive techniques for: resource allocation in datacenters, robust security defenses for machine learning algorithms, and protection of DCS/SCADA systems and critical cyber infrastructure. He also co-leads the Department of Homeland Security-sponsored DETER Network testbed project which has built the world's largest public, secure, scalable testbed for conducting next-generation cybersecurity research into worms, viruses, distributed denial of service attacks, and attacks against routing infrastructure. His principal field of interest is systems and networking: cloud computing, cybersecurity, and distributed systems.

Singer Ma (UC Santa Cruz) is a Programmer/Analyst at the UCSC Center for Biomolecular Science and Engineering. He received a B.S. in Computer Science from Harvey Mudd College in 2011 and now works on analysis of cancer data for The Cancer Genome Atlas project, focusing on somatic mutation calling.

Benedict Paten (**UC Santa Cruz**) is an assistant research scientist at the Centre for Biomolecular Science and Engineering. He did his graduate work at Cambridge University and the European Bioinformatics Institute under Ewan Birney, graduating in 2007. Since leaving Cambridge he has worked first as a postdoc with David Haussler and now as a research scientist. His work has focused mainly on algorithms for comparative genomics, in particular methods for genome alignment and evolutionary history reconstruction. He also recently helped lead a substantial community effort, dubbed the Assemblathon, for evaluating the potential of next generation sequencing for de novo assembly.

Scott Shenker (UC Berkeley) is the Carl J. Penther Professor of Engineering. He was previously Principal Scientist at Xerox PARC and then founding director of the AT&T Center for Internet Research at ICSI before joining UC Berkeley in 2004. While his past research covered such topics such as garbage collection, wireless networking, and game theory, his current research focuses on cluster computing, Internet architecture, and software-defined networking. He has been awarded the IEEE Internet Award, the ACM SIGCOMM Award, and two ACM SIGCOMM test-of-time awards, along with an honorary degree from the University of Chicago; he was also recently elected to the National Academy of Engineering. According to Microsoft Academic Search, he is the most cited author in computer science.

Taylor Sittler (UC San Francisco) is a resident in Clinical Pathology. He went to medical school at University of Massachusetts and has worked in systems biology since 2001, starting at MIT with Trey Ideker and continuing to his present-day research on the molecular classification of cancers. He received a Howard Hughes Medical Institute Medical Student Fellow Award (2004) and continuing education award (2005).

Ion Stoica (UC Berkeley) is a Professor and he received his PhD from Carnegie Mellon University in 2000. He does research on cloud computing and networked computer systems. Past work includes the Dynamic Packet State (DPS), Chord DHT, Internet Indirection Infrastructure (i3), declarative networks, replay-debugging, and multi-layer tracing in distributed systems. His current research focuses on resource management and scheduling for data centers, cluster computing frameworks, and network architectures. He is the recipient of a SIGCOMM Test of Time Award (2011), the 2007 CoNEXT Rising Star Award, a Sloan Foundation Fellowship (2003), a PECASE Award (2002), and the 2001 ACM doctoral dissertation award. In 2006, he co-founded Conviva, a startup to commercialize technologies for large-scale video distribution.

Acknowledgements

We thank Barbara Wold, Josh Stuart, Paul Spellman, Lynda Chin, Aravinda Chakravarti, Bert Vogelstein, Kenna Shaw, Tim Ley, David Bentley, Melissa Cline, Erich Weiler, Brad Smith, Steve Benz, Elaine Mardis, Matt Meyerson, Gordon Mills, Isaac Kohane, James Hamilton, Paul Berg, and David Wheeler for helpful suggestions on this manuscript. We thank Dent Earl for his assistance with graphics.